University of Alberta

EQUIVALENCE OF ACHIEVEMENT TESTS IN ENGLISH AND FRENCH DEVELOPED USING THE SIMULTANEOUS TEST DEVELOPMENT APPROACH

by

©

Jie Lin

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Measurement, Evaluation, and Cognition

Department of Educational Psychology

Edmonton, Alberta
Fall 2006

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file   Votre référence
ISBN: 978-0-494-23069-5
Our file   Notre référence
ISBN: 978-0-494-23069-5

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

ABSTRACT

The purpose of this study was to evaluate the comparability of achievement tests in English and French developed using the simultaneous test development approach. Judgmental review, statistical analysis and think-aloud interviews were conducted to address this purpose. First, five certified translators evaluated the English and French versions of each item on the Grade 9 mathematics and social studies tests for comparability in meaning and wording. Second, test performance of English-speaking and French-speaking examinees was compared in terms of psychometric characteristics, factor structures, and differential item functioning (DIF). Third, to find out whether the DIF found was related to adaptation differences, think-aloud interviews were conducted with 24 English-speaking and 39 French Immersion students. Both concurrent and retrospective verbal reports were collected. In addition, French Immersion students were asked to identify any differences in meaning and wording between the two language versions.

The English and French versions of the mathematics and social studies tests were found to be comparable in terms of meaning, psychometric characteristics, and factor structures. Larger percentage of DIF was found in social studies (42.5%) than in mathematics (19.2%). Based on the think-aloud interviews, no support for adaptation as a source of DIF was identified in any of the DIF items in mathematics or social studies. DIF in one mathematics item appears to be attributed to the interaction of a heavy load of information and inadequate French proficiency on the part of French Immersion students. For social studies, DIF in four items could be attributed to differential familiarity with key words between the two language groups, while DIF in another four items appears to be caused by differential difficulty of stimulus texts. A review of literature on French Immersion education suggests that the differential difficulty of these words/texts could be attributed to French Immersion students' lack of exposure in French outside the classroom, and inadequate proficiency in French. Last of all, limitations and implications of the study are discussed, and directions for future research are suggested.

Based on the results of this study, the simultaneous test development approach is a promising procedure for producing equally good tests across two languages.

ACKNOWLEDGEMENTS

I would like to acknowledge many people for helping me during my doctoral work. My first, and most earnest, acknowledgment must go to my advisor, Dr. W. Todd Rogers, for his expert guidance, quick response, insightful feedback, unconditional care, support and encouragement. Dr. Rogers has been instrumental in ensuring my academic, professional, financial, and moral well being in the past five years. I consider myself extremely lucky to have him as my mentor in my professional career. I am sure I could never thank him enough for all he has done to guide me through the program and thesis writing.

I am very grateful to Dr. Mark Gierl for his advice and support whenever needed during my doctoral work and dissertation writing. My thanks and appreciation also go to other members of my dissertation committee: Dr. Christina Rinaldi and Dr. Leila Ranta for generously giving their time and expertise to better my work, also Dr. Martine Cavanagh and Dr. Richard Bertrand for their thoughtful questions and comments.

This dissertation would not have been possible without the generous support from many teachers and students from Edmonton public schools. Although they are too many to name, I sincerely thank every one of them for making it possible for me to collect the interview data for this study.

My graduate studies would not have been the same without the social and academic challenges and diversities provided by all the CRAMErs. Especially, I need to express my gratitude and deep appreciation to Marilyn Abbott, Xiangming Qiu, Teresa Dawber, Shirley Li, and Antoinette Marais whose friendship, hospitality, knowledge, and

wisdom have supported, enlightened, and entertained me over the many years of our friendship.

A penultimate thank-you goes to my wonderful parents, Kuangping and Ximei, for their unwavering faith in me and unending encouragement and support. They deserve far more credit than I can ever give them.

My final, and most heartfelt acknowledgement, must go to my beloved husband, Yongkang, and daughter, Olivia. Their support, encouragement, quiet patience and unwavering love were undeniably the bedrock upon which the past eight years of my life have been built. This dissertation would not have been possible without their sacrifice of numerous hours of family time and fun. A thank-you to Yongkang for taking on the tedious task of formatting this dissertation.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION 1
Purpose of the Study 5
Definition of Terms 6
Delimitations of the Study 8
Organization of the Dissertation 8

CHAPTER II: LITERATURE REVIEW 10
Overview of Test Adaptation Methods 10
Forward/Direct Translation 10
Back Translation 11
Simultaneous Test Development 13
Differential Item Functioning (DIF) 17
Differential Item Functioning on Adapted Tests 18
Identifying Sources of DIF in Adapted Tests 24
The Use of Think-aloud Interviews for DIF Analysis 25
Overview of French Immersion Education in Canada 28
Language Proficiency 29
Effect of Language of Testing 31
Literature Summary 33

CHAPTER III: OVERVIEW OF STAGE 1
ITEM DEVELOPMENT AND PILOT TESTING 34
Item Development 34
Subject Areas and Grade Level 34
Item Writers 34
Item Writing 36
Reactions to Simultaneous Test Development 37
Item Review and Revision 39
Placement of Items in the Table of Specifications 40

Pilot Testing 40
    Pilot Test Forms 40
    Pilot Test Samples 41
    Pilot Test Results 42
Discussion 45

CHAPTER IV: METHODS AND RESULTS
JUDGMENTAL REVIEW AND ITEM SELECTION 46
Item Revision by Item Writers 47
Item Review by Certified Translators 48
    Inter-rater Agreement on Comparability of Meaning 49
    Comparability of Meaning 53
    Comparability of Wording 54
Item Revision and Selection for Field-testing 55
    Item Revision 55
    Item Selection 56

CHAPTER V: METHODS AND RESULTS
FIELD TESTING 58
Field-test Subjects 58
Descriptive Analysis 59
Structural Equivalence Analysis 61
    Non-linear Factor Analysis 61
    Multi-group Confirmatory Factor Analysis 62
Differential Item Functioning (DIF) 65
SIBTEST Results 67
    DIF: Mathematics 68
    DIF Social Studies 68

CHAPTER VI: METHODS AND RESULTS
TEACHER COMMENTS AND THINK-ALOUD INTERVIEWS 70

Teacher Comments ... 70
  Mathematics ... 70
  Social Studies ... 72
Think-aloud Interviews ... 74
  Interview Sample ... 75
  Instruments ... 76
  Interviewers ... 77
  Interview Procedure ... 78
  Data Analysis ... 78
  Inter-coder Agreement ... 80
Interview Results for Mathematics ... 81
  DIF Items ... 82
  Non-DIF Items ... 85
Interview Results for Social Studies ... 85
  DIF Items ... 85
  Non-DIF Items ... 98

CHAPTER VII: DISCUSSIONS AND CONCLUSIONS ... 99
Summary of Research Questions ... 99
Discussion ... 100
  Research Question 1 ... 100
  Research Question 2 ... 103
  Research Question 3 ... 104
  Implications for Test Administration to French Immersion Students ... 109
Efficacy of Simultaneous Test Development Approach ... 109
  Efficiency ... 109
  Effectiveness ... 111
  Implications for practice ... 113
Limitations and Directions for Future Research ... 114

Bibliography ... 116

LIST OF TABLES

Table 1: Language of Testing Study by Alberta Education (1990-1992) 32

Table 2: Background of Item Writers 37

Table 3: Pilot Test Sample Sizes 42

Table 4: Distribution of Items by Class 43

Table 5: Mathematics Translator Review Results 51

Table 6: Social Studies Translator Review Results 52

Table 7: Comparability of Meaning for Mathematics and Social Studies Items 53

Table 8: Psychometric Characteristics of the Mathematics and Social Studies Tests...... 60

Table 9: NOHARM Fit Indices for 1- and 2-Dimensional Models 62

Table 10: Tests for Model Equivalence between English-speaking and French Immersion Examinees 65

Table 11: SIBTEST Results for the Mathematics and Social Studies Tests 67

Table 12: Distribution of DIF in the Mathematics Tests 68

Table 13: Distribution of DIF in the Social Studies Tests 69

Table 14: Teacher Comments on the Mathematics Tests 71

Table 15: Teacher Comments on the Social Studies Tests 73

Table 16: Inter-rater Agreement for the Coding of Interview Data 81

Table 17: Rating of the Equivalence of Mathematics Items by French Immersion Students 82

Table 18: Rating of the Equivalence of Social Studies Items by French Immersion Students 86

Table 19: Comparison of Item Development Time 111

FIGURE

Figure 1: Model for the Concurrent Development 15

LIST OF APPENDICES

Appendix A: Grade 9 Mathematics Subject Blueprint 126
Appendix B: Grade 9 Social Studies Subject Blueprint 127
Appendix C: Mathematics Translator Review Results 128
Appendix D: Social Studies Translator Review Results 129
Appendix E: Mathematics Achievement Test (English) 130
Appendix F: Mathematics Achievement Test (French) 143
Appendix G: Social Studies Achievement Test (English) 156
Appendix H: Social Studies Achievement Test (French) 177
Appendix I: Parent Consent Letter 198
Appendix J: Verbal Report Instructions for Social Studies 200
Appendix K: Observation Sheet 202
Appendix L: Confidentiality Agreement 203
Appendix M: Coding Scheme for Mathematics 204
Appendix N: Coding scheme for Social Studies 205
Appendix O: Coding Sheet for English Transcripts 206
Appendix P: Coding Sheet for French Transcripts 207

CHAPTER ONE

INTRODUCTION

The adaptation of educational tests is becoming ever more important due to the marked increase in international, national, and state/provincial testing in a time when an increasing number of students are studying in different languages. The Third International Mathematics and Science Study (TIMSS), for example, involved over 45 countries and 30 languages. Fifty-seven countries are participating in the Program for International Student Assessment (PISA) in 2006. In Canada, two official languages are used at the federal level. In keeping with this requirement, the provincial ministries of education introduced French Immersion programs to allow students to acquire competence in French while learning the content as set out in provincial programs of study. At the same time, the provincial testing programs in the provinces with such tests expanded the testing program to include testing in French as well as in English. In spite of the expectation that the original tests and the subsequent adaptations are equivalent in terms of the constructs they measure, research has shown otherwise (e.g., Allalouf, Hambleton, & Sireci, 1999; Angoff & Cook, 1988; Budgell, Raju, & Quartetti, 1995; Ercikan 1998, 1999; Gierl, 2000; Gierl, Rogers, & Klinger, 1999; Hambleton, 1993; Sireci & Berberoğlu, 2000; Sireci, Fitzgerald, & Xing, 1998; Solano-Flores, Trumbull, & Nelson-Barber, 2002; Tanzer, 2005; van der Vijver & Tanzer, 1997).

Before proceeding, a note needs to be made about the distinction between *test adaptation* and *test translation*. The term *test adaptation* is generally preferred to *test translation* in the field of cross-cultural studies. According to Hambleton (2005),

> *Test adaptation* is broader and more reflective of what should happen in the process of preparing a test constructed in one language and culture for use in another language and culture. Test adaptation includes all the activities from deciding whether or not a test could measure the same construct in a different language and culture, to selecting translators, to deciding on appropriate accommodations to be made in preparing a test for use in a second language, to adapting the test and checking its equivalence in the adopted form. Test

> translation is only one of the steps in the process of test adaptation and even at this step, adaptation is often a more suitable term than translation to describe the actual process that takes place. This is because translators are trying to find concepts, words, and expressions that are culturally, psychologically, and linguistically equivalent in a second language and culture, and so clearly the task goes well beyond simply preparing a literal translation of the test content. (p. 4)

That is to say, the test adaptation process involves more than literal word-to-word translation, allowing for more complex adaptations in order to ensure the equivalence of the tests across languages. Therefore, *test adaptation* is used instead of *test translation* in this dissertation except for some well-accepted terminologies, such as forward translation and back translation.

A number of factors have been identified that contribute to the increased interest in test adaptation: (a) the need to enhance the fairness of comparison of individuals and groups from different language and cultural backgrounds, (b) the increased need for comparative studies across national, ethnic, and cultural groups, and (c) the increased need for comparison of student achievement across the world (Hambleton, 1994; Sireci, 1997). Hambleton (2005) presented a good example of poor test adaptation that illustrates why there is growing interest in how test adaptations are executed. In an international comparative study of reading proficiency, American students were asked to compare pairs of words and identify them as similar or different in meaning. For the pair "sanguine – pessimistic", only 54% of American students answered correctly. In the top-performing non-English-Speaking country, however, 98% of the students answered this question right. It was later discovered that there is no equivalent for "sanguine" in their language. Consequently, the counterpart "optimistic" was used, which in turn made the question easier. This example provides a good illustration of how adaptation can affect the validity of test score interpretation in multilingual assessments.

The two most popular designs for adapting tests are forward translation and back translation. With the forward-translation design, which is sometimes called direct translation, one or more translators translate a test from the source language to the target language. Then the equivalence of the two versions of the test is checked by another group of translators. Revisions can then be made based on the recommendations of the second group of translators (Geisinger, 1994; Hambleton, 2005). The main advantage of the forward-translation design is that direct judgments can be made about the equivalence between the two versions of the test (Hambleton, 2005). Further, forward translation generally involves less time compared with back translation. However, the weaknesses of the forward-translation design reside mainly in (a) the high level of inference that must be made by the translators about the equivalence of the two versions of the test, and (b) the inability of monolingual test developers and researchers to judge test equivalence (Hambleton, 2005). Although it may continue to be one of the most frequently used methods for test adaptation, "direct translation from the source language to the target language has been repudiated as an unreliable method for achieving language equivalence" (Brislin, 1970; Sperber, Devellis, & Boehlecke, 1994; Triandis, 1976; Werner & Campell, 1970 in Erkut, Alarcon, Gracia Coll, Tropp, & Vazquez Garcia, 1999, p. 208).

The back-translation design is the best known and most frequently applied procedure for adapting tests. A back-translation design involves (a) forward translation of a test into the target language, and (b) back translation of the translated test into the source language to monitor retention of the original meaning in the source language (Behling & Law, 2000; Hambleton & Bollwark, 1991). To the extent that the original and back-translated versions of the test in the source language are similar, evidence is provided for the equivalence of the original and translated tests. The back-translation design enables researchers who are not fluent in the target language to evaluate the quality of translation by comparing the original and back-translated source language tests (Gierl et al., 1999; Hambleton, 2005). Researchers generally agree that the back-translation design provides an overall check on adaptation quality and can be used to detect adaptation differences (Ellis, 1989; Hambleton, 1993, 2005; van de Vijer & Leung, 1997).

Despite the advantages, concerns have been raised with the back-translation design. First, Stansfield and Kahl (1998) contended that the differences between the original and adapted tests might be due to problems with the back translation and not to problems with the forward translation. The back translation is just as likely to contain translation errors as is the forward translation. Basically, "one is left with two translations with no verification of the quality of either" (Stansfield & Kahl, 1998, p. 6). Further, in the process of back translation skilled translators may improve the test when the original translation is poor (Hambleton, 1993). Without direct evaluation of the source-to-target translation, one can never be certain whether the discrepancies between the original and back-translated tests in the source language are attributed to problems in the forward translation, the back translation, or both. Second, the back-translation design may result in literal translation at the expense of connotation, naturalness, and comprehensibility across languages, especially when the forward translators are aware that there will be back translation (Stansfield & Kahl, 1998; van de Vijer & Leung, 1997). Above all, a common weakness of both forward- and back-translation designs is that monolingual developers' lack of competence in other languages or cultures may lead to ethnocentrism and linguistic or cultural specifics in the source test that make it almost impossible to create equally "good" test versions in the target language (Rogers, Gierl, Tardif, Lin, & Rinaldi, 2003).

In response to the above concerns, Tanzer (2005) called for simultaneous development as an alternative to ensure cross-lingual/cross-cultural validity. In simultaneous test development, the test is explicitly created for use in a multilingual/multicultural assessment. When bilingual tests are developed simultaneously, bilingual/bicultural test writers develop the source and target forms at the same time. The two language forms are equally open to modification in the process of test development. Consequently, language and culture specifics can be detected and removed at the early stages of test development, thereby reducing the risk of construct bias and maximizing linguistic and cultural decentering (Solano-Flores et al., 2002; Tanzer, 2005). The potential advantage of simultaneous test development is ensuring that the quality of the test is equally good across languages.

To date, only a few studies have investigated the utility of variants of the simultaneous test development approach. Erkut et al. (1999) proposed a dual-focus approach to creating bilingual measures. The concept-driven approach implemented by bilingual/bicultural experts was reported to be effective in minimizing the chances of obtaining non-equivalent test forms due to lack of correspondence in concepts and lack of equivalence in item wording. This approach was also instrumental in maximizing conceptual and linguistic equivalence of the two versions of the measure. Solano-Flores et al. (2002) documented their task development process and concluded that concurrent test development "allows assessment developers to generate high-quality assessments for linguistic minorities by supporting them to give deeper consideration to subtle language issues and culture as part of their discussion throughout the entire process of assessment development" (p. 127). Unfortunately, neither study provided convincing empirical (e.g., DIF) or substantive (e.g., data from external judges or students) evidence for the degree of equivalence of the bilingual tests developed. More research is needed to determine whether the hypothesized advantages of the simultaneous test development approach are indeed tenable with reasonable effort and cost.

Purpose of the Study

This study is part of a large-scale research project designed to investigate the validity and utility of the simultaneous approach to the development of equivalent achievement tests in French and English. The major objectives of the large-scale project were to a) develop Grade 9 mathematics and social studies tests in French and English employing the simultaneous approach, b) validate the tests produced, and c) evaluate the utility of the simultaneous approach in terms of cost-effectiveness, ease of implementation, and quality of tests with regard to degree of biasedness and presence of measurement error (Rogers et al., 2003). The first stage of the research project was reported in Rogers et al. (2003). In this stage, six bilingual item writers, three for mathematics and three for social studies, were recruited to develop the initial French and English versions for each item at the same time. They wrote each item in one language and then immediately translated it into the second language. They were not allowed to move on to the next item until they had made sure that the items in both languages meant the same and called for the same level of thinking for the target students. After the item writers reviewed each other's work, the retained items were pilot tested. The pilot test results revealed that the French-speaking examinees outperformed the English-speaking examinees in both mathematics and social studies. What was not clear is why this difference occurred. Possible reasons include non-equivalence of the tests constructed in the two languages, the presence of socio-economic differences between the two language groups (i.e., real differences in ability), or a combination of both (Rogers et al., 2003).

Consequently, the purpose of the present study was to disentangle these two issues to obtain a clearer view of the efficacy of the simultaneous approach in reducing construct bias and enhancing linguistic and cultural decentering. In particular, the following research questions were investigated:

1. How comparable are the English and French versions of the Grade 9 achievement tests in mathematics and social studies constructed using the simultaneous test development approach?
2. Is there evidence of differential item performance for English- and French-speaking examinees on the above tests?
3. If so, to what degree is the source of differential item performance related to adaptation differences?

Question 1 was addressed employing evidence from judgmental review, item analysis and factor analysis. Question 2 was addressed by the analysis of differential item functioning. Teacher comments and students' interview data were used to address Question 3. Unlike previous applications of the simultaneous test development (e.g., Erkut et al., 1999; Solano-Flores et al., 2002), this study utilized comprehensive data from both empirical and substantive sources to evaluate the equivalence of the bilingual tests developed.

Definition of Terms

Test adaptation: Test adaptation includes all the activities from deciding whether or not a test or a test item could measure the same construct in a different language and culture, to selecting translators, to deciding on appropriate accommodations to be made in preparing a test or an item for use in a second language, to adapting the test and checking the equivalence between the language forms (Hambleton, 2005, p. 4).

Forward translation: Forward translation, or direct translation, involves (a) translation of a test from the source language to the target language by one or more translators, (b) evaluation of the equivalence of the two versions of the test by another group of translators, and (c) revisions based on the recommendations of the second group of translators (Geisinger, 1994; Hambleton, 2005).

Back translation: Back translation involves (a) translation of a test from the source language to the target language, (b) independent translation of the translated test back into the source language, and (c) comparison of the two versions of the test in the source language until discrepancies in meaning are resolved or adjusted (Brislin, 1970, 1986).

Decentering: Decentering is a translation process in which the source and the target language versions are equally important and open to modification (Brislin, 1973, pp. 37-38).

Simultaneous test development: The simultaneous test development approach involves developing a new test for use in a number of predefined cultural groups and/or languages. That is, two or more language versions of a test are developed simultaneously. A committee approach is usually employed that involves a multilingual task force from various cultural backgrounds and with complementary expertise in mainstream psychology (including knowledge of the construct and its measurement), psychometrics, test construction techniques as well as cultural psychology, cross-cultural psychology, and linguistics (Tanzer, 2005).

Differential Item Functioning (DIF): DIF occurs when examinees from different groups have different probabilities or likelihood of success on a item after conditioning or matching on the ability the test is intended to measure (Shepard, Camilli, & Averill, 1981).

Item bias: Item bias refers to DIF that is attributed to systematic error in how a test item measures a construct for members of a particular group (Camilli & Shepard, 1994).

Item impact: Item impact refers to DIF that is attributed to group discrepancy in item performance that reflects actual knowledge and experience differences on the construct of interest (Gierl et al., 1999).

English-speaking students: English-speaking students refer to students who receive instruction in English. English-speaking students in Alberta account for about

95% of the Grade 9 student population (Jolanta Wojcik, personal communication, June 8, 2005).

French-speaking students: French-speaking students refer to students who are instructed mostly in French. In Alberta, French-speaking students account for about 5% of Grade 9 student population. Among them, 90% are French Immersion students and 10% are Francophone students (Jolanta Wojcik, personal communication, June 8, 2005).

French Immersion students: French Immersion students are students for whom French is not a first language and who are attending French Immersion programs to become functionally fluent in French. The language of instruction is French in many subject areas.

Francophone students: Francophone students are students for whom French is their first language. The majority attend French schools where the language of instruction is in French. French schools are designed for students with at least one French-speaking parent. Students in French schools are expected to master French as a mother tongue, and establish a sense of identity and belonging to the French community.

Delimitations of the Study

This study focused mainly on the degree of equivalence of the achievement tests in English and French developed using the simultaneous approach, the degree that the items function differentially between the two language groups, and the degree to which DIF between the two language versions may be attributed to adaptation-related differences. Other related issues such as the effects of bilingualism, cultural differences, and curricular differences were beyond the scope of the study.

Organization of the Dissertation

This dissertation is organized in seven chapters. Chapter 1 serves as an introduction to the study. Chapter 2 contains a review of the literature, including (a) an overview of two well-known test adaptation methods and the simultaneous approach, (b) a summary of research on the identification of sources of differential item functioning

(DIF) on adapted tests, (c) a discussion of the use of think-aloud interviews in identifying sources of DIF, and (d) an outline of French Immersion education in Canada. Chapter 3 is a summary of Stage I, which provides the background information for the present study. Given the sequential nature of the procedures used in this study, both the methods used and the results obtained using these methods are presented together in the next three chapters. Chapter 4 presents the methods and results for the item review by certified translators, and the selection of items for the final test forms. The methods and results for statistical analysis of the field test data are provided in Chapter 5. Chapter 6 describes data collection procedures and results of analysis of teacher comments and think-aloud protocols of the English-speaking and French Immersion students. Chapter 7 contains a summary of the procedures and findings of the study, followed by a discussion of limitations and implications for practice and future research.

CHAPTER TWO

LITERATURE REVIEW

The literature review is organized in five sections. In the first section, the literature relevant to test adaptation methods is reviewed. The second section covers critical reviews of research pertaining to differential item functioning on adapted tests, especially the sources of DIF on adapted tests. The third section contains a brief discussion of literature on the use of think-aloud interviews for DIF analysis. The fourth section offers an overview of French Immersion education in Canada. The last section contains a summary of the main conclusions drawn from the literature.

Overview of Test Adaptation Methods

The two most popular designs for adapting tests are forward translation and back translation.

Forward/Direct Translation

With the forward-translation design, one or more translators translate the test from the source language to the target language. Next, the equivalence of the two versions of the test is checked by another group of translators. Revisions can then be made based on the recommendations of the second group of translators (Geisinger, 1994; Hambleton, 2005). The main advantage of the forward-translation design is that direct judgments can be made about the equivalence between the two versions of the test (Hambleton, 2005). The technique is practical and generally involves less time compared with back translation. However, the weaknesses of the forward-translation design include (a) the high level of inference that must be made by the translators about the equivalence of the two version of the test and (b) the inability of the monolingual test developers and researchers to judge test equivalence themselves (Hambleton, 2005). As Behling and Law (2000) noted, a well-documented problem in the forward-translation design involves the difficulty in obtaining a truly representative consensus through group discussion. Therefore, it can never be guaranteed that the revision group will arrive at the best possible translation. Although it continues to be one of the most frequently used methods for test adaptation, "direct translation from the source language to the target language has been repudiated as an unreliable method for achieving language equivalence" (Brislin, 1970; Sperber et al., 1994; Triandis, 1976; Werner & Campell, 1970 in Erkut et al., 1999, p. 208).

Back Translation

Back translation is a well-known and commonly used technique for adapting tests. According to Behling and Law (2000), the back-translation design is an iterative process in which each cycle involves four steps:

1. A bilingual individual translates the source language instrument into the target language.
2. A second bilingual individual with no knowledge of the wording of the original source language document translates this draft target language rendering back into the source language.
3. The original and back-translated source language versions are compared.
4. If substantial differences exist between the two source language documents, another target language draft is prepared containing modifications designed to eliminate the discrepancies. (p. 20)

The back-translation process is repeated until the two source language tests are identical or close to identical. To the extent that the original and back-translated versions of the test in the source language are similar, evidence is provided for the equivalence of the original and translated tests.

The state-of-the-art method for developing bilingual measures, according to Erkut et al. (1999), has been back translation used in combination with the "decentering" technique (Brislin, 1970; Werner & Campell, 1970). Decentering is "a translation process in which the source and the target language versions are equally important and open to modification" (Brislin, 1973, pp. 37-38). Typically, following forward translation and backward translation, if substantial discrepancies are found between the two source language versions of a test, the target language and source language versions are equally open to change. In some cases, the back-translated version is simply substituted for the original source language item.

Generally speaking, the back-translation design enables researchers who are not fluent in the target language to evaluate the quality of translation by comparing the original and back-translated source language tests (Gierl et al., 1999; Hambleton, 2005). Researchers generally agree that the back-translation design provides an overall check on the quality of the adaptation and can be used to detect adaptation differences (Ellis, 1989; Hambleton, 1993, 2005; van de Vijer & Leung, 1997).

Despite the advantages, concerns have been raised about back translation. For example, back translation is conducted in order to identify problems in forward translation, but differences between the original and adapted tests might be due to problems with the back translation and not to problems with the forward translation. The back translation is just as likely to contain translation errors as is the forward translation. Basically, "one is left with two translations with no verification of the quality of either" (Stanfield & Kahl, 1998, p. 6). At least four factors can lead to the resemblance of the original and back-translated source language versions while in fact the forward translation is poor (Behling & Law, 2000). First, the forward and backward translators may share a set of rules for translating certain non-equivalent words and phrases (e. g., *amigo* in Spanish and *friend* in English are not always equivalent). Second, back-translators may be skilled enough to make sense of a target language version even if it depicts the original version poorly, and thus achieve a back-translated version that is misleadingly close to the original source language version. Third, the target version of the test may retain inappropriate aspects of the source language test such as the same grammatical structure and spelling, facilitating back translation but hiding serious shortcomings in the target language test because the two source language tests would appear similar (Brislin, 1970; Sperber et al, 1994). Last, Hambleton (1993), as cited in Gersinger (1994) and Behling and Law (2000), suggested that "when translators knew that their work was going to be subjected to back translation, they would use wording that ensures that a second translation would faithfully reproduce the original version rather than a translation using the optimal wording in the target language" (Geisinger, 1994, p. 306). Without a direct evaluation of the source-to-target translation, one can never be certain whether the discrepancies between the original and back-translated tests in the source language are attributed to problems in the forward translation, the back translation, or both. Hambleton (2005) has concluded that although the back-translation design can "identify problems in a test adaptation process, it would rarely provide a sufficient amount of evidence to support the valid use of an adapted test" (p. 13).

Above all, Brislin (1986) asserted that even when the original and translated tests are linguistically equivalent, they may not be psychologically equivalent. The translated version of a test may not capture entirely the thinking associated with the source language and culture (Greenfield, 1997). A common weakness of both forward- and back-translation designs is that monolingual developers' lack of competence in other languages or cultures may lead to ethnocentrism and linguistic or cultural specifics in the source test that make it almost impossible to create equally "good" test versions in the target language (Rogers et al., 2003). That is, monolingual/monocultural test developers of the original test are usually experts in the subject matter as well as in the source language and culture, but they may not be equally knowledgeable in the target languages and cultures. As a result, the forward translators may find it difficult, if not impossible, to create a test in the target language that is equivalent to the original test in terms of linguistic, cultural, and psychological perspectives.

Simultaneous Test Development

The earliest form of the simultaneous test development approach can be traced back to the 1970s, when Werner and Campbell (1970) recognized the problems arising from having a source language and a target language. They proposed that tests/measures should be developed jointly in two cultures using the decentering method. The team of researchers need to have expertise in the subject matter to be assessed as well as the two cultures. However, applications of the simultaneous test development have been rare.

Drawing on Werner and Campbell (1970) and Triandis (1976), Erkut et al. (1999) proposed a dual-focus approach, a variant of the simultaneous approach. The dual-focus approach requires a bilingual/bicultural research team and employs a concept-driven rather than translation-driven approach to attaining conceptual and linguistic equivalence between the two language versions of a measure. Five steps are required to implement the dual-focus approach: (a) collaboration of a bilingual/bicultural research team in defining research questions, study design, and implementation; (b) operationalization of the content area of the construct(s) by selecting concepts that provide equally valid definition of the constructs in both cultures; (c) generation of items to measure the common concepts with special attention paid to the wording so as to ensure the same level of difficulty, affect, and clarity of meaning; (d) evaluation of the items by bilingual and monolingual focus groups (members of the communities for whom the measure is intended); and (e) estimation of psychometric characteristics of the measure in regard to the validity and reliability of the two language versions. Erkut et al. illustrated that when working from a common conceptual base, both languages become the target languages: thus, the dual focus. Success in the application of the dual-focus approach was reported in the development of the Psychological Acculturation Scale. When 36 self-identified bilinguals were administered both the English and Spanish versions in a counterbalanced random order, the correlation between their scores on the two versions was $r = 0.94$. Nevertheless, while $r = 0.94$, the mean absolute difference (MAD) could be greater than zero (e.g., when their scores on one version of the test are different than their scores on the other version). That is, the correlation index alone is not enough to tell how closely the students scored on the two versions of the test. Besides, no further evidence in regard to the equivalence of the two versions of the test was provided, such as DIF analysis, external judgmental review, or student input.

Some advantages of the dual-focus approach were noted by the authors. First, the concept-driven approach implemented by bilingual/bicultural experts minimizes the chances of obtaining non-equivalent test forms due to lack of correspondence in concepts and lack of equivalence in item wording. Second, the dual-focus approach is time-efficient in establishing the linguistic equivalence of the bilingual measures. "The simultaneous development and examination of items in both languages by experts in the subject matter and cultures bypass the lengthy process of back translation" (p. 216). Third, the dual-focus approach can minimize translation errors that may result from employing translators who do not have knowledge of the subject matter, and thereby maximize conceptual and linguistic equivalence of the two versions of the measure. In regard to the limitations of the dual-approach method, the authors remarked that the method is most straight-forward when only two languages are involved. When multi-language/cultures are involved, the task of finding researchers indigenous to all the languages and cultures can be challenging.

Similarly, Solano-Flores et al. (2002) proposed and implemented what they called the concurrent development of dual language assessments, another variant of the simultaneous approach. Figure 1 illustrates the process of implementing the concurrent assessment development model for two language versions of the same test: mainstream language (Version A) and minority language (Version Alpha). The model is based on the use of "shells" or blueprints that specify the structural and formal characteristics of the items. The shells provide the test developers with directions for generating items of similar structures and complexities (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1999). After two teams of test developers create a draft for each language independently, both versions go through the same process of review- tryout-revision iterations, and both versions are equally open to modification. According to the authors, "concurrent" emphasizes the fact that "the two languages converge or interact throughout the entire process of assessment development" (p. 111).

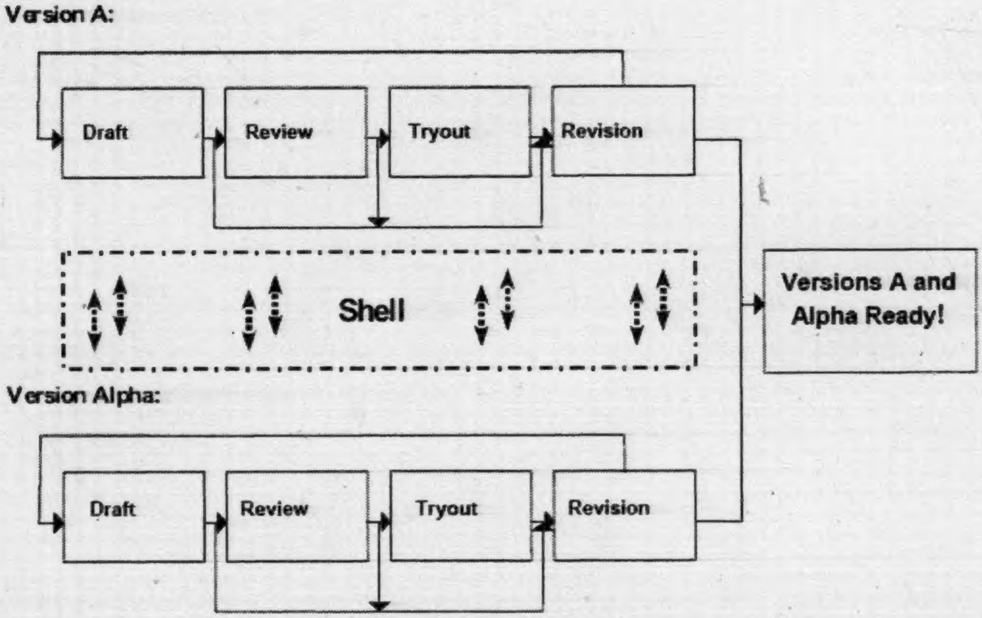


Figure 1. Model for the concurrent development of two language versions (A and Alpha) of the same assessment (Solano-Flores et al., 2002, p. 110).

The concurrent development model was used by Solano-Flores et al. (2002) to develop a set of constructed-response mathematics exercises for Grade 4 in English and Spanish. Seven experienced bilingual teachers were divided into two teams: the Spanish team (4 teachers) and the English team (3 teachers). To ensure comparable structures and appearance of the items in the two languages, the shells were created by the researchers and improved by the teachers. Based on the shells, the English and Spanish teams worked independently developing the items. They then met for a group discussion to make sure that the meaning, wording, and complexity of the items were comparable. Changes on the shell regarding format and directions asking students to provide a response were made to ensure that the items were appropriate to the characteristics of both groups of examinees.

Solano-Flores et al. (2002) claimed that one fact spoke to the efficiency of the concurrent development model. The training of the teachers to generate items using the shell took no longer than three hours. There was no evidence that using the concurrent development model would mean longer development time or higher costs than traditional test development methods. An indicator of the effectiveness of the model is that it actually allowed teachers to pay attention to language and culture in their discussions throughout the process of test development. By the third session teachers were observed to speak with greater depth and specificity about linguistic and cultural issues. The authors concluded that the concurrent development model enabled test developers to generate high-quality tests for linguistic minorities by encouraging them to give deeper consideration to subtle language and culture issues throughout the test development process. Although Solano-Flores et al. (2002) provided preliminary evidence for the efficacy of the concurrent development model by examining the test development process, further evidence such as the empirical and substantive comparability of the final two language versions of the test based on student responses to the two versions, is needed before a fair evaluation of the concurrent model can be achieved.

More recently, Tanzer (2005) called for simultaneous development in the newest publication on test adaptation, *Adapting Educational and Psychological Tests for Cross-cultural Assessment* (Hambleton, Merenda, & Spielberger, 2005). Tanzer first illustrated problems often encountered in multilingual test applications using successive approaches (e.g., forward translation and back translation), including problems caused by single items, culturally incompatible test designs, and ethnocentric instructions and administration procedures. Given the limitations of the successive approaches, Tanzer recommended the simultaneous test development approach. In the simultaneous approach, a new test is developed for use in a number of predefined cultural and/or language groups. It usually involves a "committee approach", which is a multilingual team from various cultural backgrounds with complementary expertise in psychology, psychometrics, and test construction as well as cultural and cross-cultural psychology and linguistics. With the simultaneous approach, language and culture specifics can be detected and removed at the early stages of test development, thereby reducing the risk of construct bias and maximizing linguistic and cultural decentering (Solano-Flores et al., 2002; Tanzer, 2005). The potential advantage of simultaneous test development is ensuring that the quality of the test is equally good across languages. What is challenging with the simultaneous approach is that the assembly team of test developers with expertise in psychometrics, cross-cultural psychology, and linguistics may be hard to find, especially when more than two languages or cultures are involved.

Although the idea of simultaneous test development has been around for over three decades, until 2002 only a few studies regarding the application of this method had been published (e.g., Erkut et al., 1999; Solano-Flores et al., 2002). Unfortunately, none of these studies provided convincing empirical (e.g., DIF) or substantive (e.g., feedback from external reviewers or students) evidence for the degree of equivalence of the bilingual tests developed. More research is needed to determine whether the hypothesized advantages of the simultaneous test development approach are indeed tenable.

Differential Item Functioning (DIF)

Differential item functioning (DIF) analysis is a procedure used to identify items that function differently between different groups after controlling for ability. It is based on the assumption that test takers who have the same knowledge (based on total test scores) should perform in similar ways on individual test items regardless of their gender, race, or ethnicity. DIF occurs when examinees from different groups have different probabilities or likelihood of success on an item after matching on the ability the test is intended to measure (Shepard, Camilli, & Averill, 1981). Once the DIF items are detected statistically, there is a need for substantive interpretation to determine whether the items display bias or impact. *Item bias* is generally defined as "invalidity or systematic error in how a test item measures [a construct] for members of a particular group" (Camilli & Shepard, 1994, p. 8). *Item impact* refers to "group discrepancy in item performance that reflects actual knowledge and experience differences on the construct of interest" (Gierl et al., 1999, p. 355). If an item is biased, it should be either deleted or revised. If an item demonstrates impact, it should be retained but further investigation may be necessary to explore why one group scored higher than another group.

Differential Item Functioning on Adapted Tests

The comparability of test items across different language groups is often evaluated using DIF analysis. The presence of DIF in an item suggests that examinees with the same overall ability score who belong to different language groups have different success rates on this item. Large proportions of DIF items may seriously weaken the equivalence of the original and adapted tests and raise questions concerning the validity of any score interpretations derived from the adapted tests.

Studies on the psychometric characteristics of adapted tests have revealed that the percentage of DIF items on some tests is large. For example, Ercikan (1999) found that 58 out of 140 science items (41.4%) from the Third International Mathematics and Science Study (TIMSS) exhibited DIF when Canadian English- and French-speaking examinees were compared. Gierl et al. (1999) reported that 26 of 49 social studies items (53.1%) on a Canadian Grade 6 achievement test adapted from English to French displayed DIF. More recently, Ercikan and Koh (2005) reported that as many as 110 out of 139 science items (79.1%) on the TIMSS displayed DIF when American and French examinees were compared.

Although a number of researchers have looked at how examinees writing different language versions of a test perform differentially at the item level, a limited number of studies have attempted to investigate the sources of DIF on these adapted tests employing substantive evidence. These studies are discussed below.

Ercikan (1998) explored the effect of adaptation on the comparability of test items in the International Association for the Evaluation of Educational Achievement (IEA) science tests. Assessment data on 5,543 English- and 2,348 French-speaking students in Canada were examined for DIF using the Mantel-Haenszel procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959). Eighteen out of 70 items (25.7%) were flagged as DIF. Next, with the help of several translators, Ercikan reviewed the original English version, the translated French version, and the back-translated English version of the tests. Eight of the 18 DIF items (44.4%) were attributed to adaptation-related differences. Three adaptation-related problems were identified: "a) differential frequency, difficulty or commonness of vocabulary; b) differential length or complexity of sentences; and c) differential contextual meaning of vocabulary" (Ercikan, 1998, p. 552). Besides, Ercikan also noted the presence of stronger DIF in favor of the group who took the original form of the tests.

Gierl et al. (1999) examined adaptation-related DIF using both statistical and judgmental methods. Response data from a random sample of 2,200 English- and 2,200 French-speaking examinees were compared on Grade 6 mathematics and social studies provincial achievement tests in Alberta, Canada. Three statistical procedures were applied to identify DIF--Mantel-Haenszel, SIBTEST (Simultaneous Item Bias Test; Shealy & Stout, 1993), and logistic regression (Swaminathan & Rogers, 1990). The DIF items flagged by the three methods were relatively consistent, but not identical, with Mantel-Haenszel being the most conservative and logistic regression being the most liberal. Seven out of 50 items (14.0%) in mathematics and 26 out of 49 items (53.1%) in social studies were identified to exhibit DIF by at least two of the procedures. For the judgmental analysis, two certified translators independently back-translated the achievement tests from French to English. Without knowledge of the DIF status of the items, three reviewers independently evaluated the comparability of each item using the original English form, the translated French form, and the two back-translated forms. Following group discussions, the reviewers compared the statistical outcomes with their ratings, and adjusted their ratings where necessary. The final rating of the degree of item comparability was created based on the agreement of at least two out of three reviewers. Two of the seven DIF items (28.6%) in mathematics were attributed to adaptation-related differences, while seven out of the 26 DIF items (26.9%) in social studies were identified to be related to adaptation. The authors concluded that the discrepancy between statistical and judgmental results in social studies might be attributed to the inflated Type I error because of the use of an inadequate conditioning variable. When the proportion of DIF items is large, total test score (or a latent version of total test score as with the SIBTEST procedure) may not be a valid variable to match the examinees. Iterative purification can be an option, but it was unknown how the removal of large number of items would affect the construct and content representation on the tests.

Allallouf et al. (1999) investigated DIF in relation to item type and causes of DIF in adapted items contained in the verbal subtest of Psychometric Entrance Test (PET) in Israel. Data from large samples (ranging from 1,485 to 7,150) of Hebrew- and Russian-speaking examinees on three test forms were analyzed for DIF using the Mantel-Haenszel procedure. Of the 125 items considered across the three forms, 42 (33.6%) were identified with DIF, with the greatest incidence of DIF on the analogy items. Five translators were then recruited to independently evaluate 60 items (42 DIF and 18 non-DIF) without knowledge of the DIF classification. Specifically, they were asked to predict item characteristics such as whether the item displayed DIF, the direction and magnitude of DIF, and the reason for DIF. Following meetings with the three researchers where statistical information on DIF was presented, consensus was reached regarding the causes of DIF for 35 out of the 42 DIF items (83.3%). The causes identified included: differences in difficulty of vocabulary, differences in content, differences in format, and differences in cultural relevance. Allalouf et al. realized that their study might be limited because only two language groups were used. They argued that replication using different languages and multiple forms would help to ensure the generalizability of the findings across language groups.

Gierl and Khaliq (2001) employed a confirmatory analysis to identify sources of differential item and bundle functioning on adapted tests. They used data from 3,000 English-speaking and 2,115 French Immersion students on the 1997 Grade 6 mathematics and social studies provincial achievement tests in Alberta, Canada. Similarly, data was also taken from 3,000 English-speaking and 2,115 French Immersion students on 1997 Grade 9 mathematics and social studies provincial achievement tests. This study followed a DIF analysis paradigm proposed by Roussos and Stout (1996a), which combined substantive and statistical analyses by linking both to the Shealy-Stout (1993) multidimensional model of DIF. First, a substantive analysis was conducted to generate DIF hypotheses. For this purpose, an 11-member review committee was formed, including six bilingual reviewers (three translators, one editor, and two test developers) and five monolingual English-speaking reviewers (two psychometricians, one test developer, and two directors for test development). The committee members reviewed all the DIF items identified using SIBTEST from 1996 Grade 6 mathematics and social studies provincial achievement tests (Gierl et al., 1999). Without knowledge of the statistical outcomes of the DIF items, the committee members were asked, for each item, to identify any translation problems or differences, to depict the sources of the translation differences, and to specify which group this item would favour. As a result, four sources of adaptation-related DIF were identified. "(a) Source 1: omissions or additions that affect meaning, (b) Source 2: differences in words, expressions, or sentence structure of items that are *inherent* to the language and/or culture, (c) Source 3: differences in words, expressions, or sentence structure of items that are *not inherent* to the language and/or culture, and (d) Source 4: differences in item format."(p. 173). Gierl and Khaliq (2001) noted that the sources of translation DIF identified by them and the sources identified by Allalouf et al. (1999) were quite similar, with some minor differences in classification. For example, unlike Allalouf et al. (1999), Gierl and Khaliq (2001) made a distinction between differences in words, expressions, and sentence structure that *are inherent* to language and/or culture and those that are *not inherent* to language and/or culture.

Second, to validate the sources of translation DIF identified in the first stage, two certified translators used the four-source framework to classify DIF items in the 1997 Grade 6 and 9 mathematics and social studies achievement tests. The two translators independently sorted the items into eight categories--four sources of translation DIF across two language groups-- for each subject in both grade levels. Finally, each of the eight categories of items were tested separately as a bundle against a purified matching subtest (with DIF items removed) using SIBTEST. In terms of the comparability between the translators' predictions and statistical outcomes, the results were not consistent across the subjects, with better agreement in mathematics than in social studies. Seven of the eight bundles (87.5%) created by translators in mathematics produced significant results, while only eight of the 13 bundles (61.5%) created in social studies were significant. Across subjects and grade levels, Source 3 bundles contained the largest number of DIF items and were consistently identified. The Source 4 bundle, in contrast, was found only once. The items predicted to be associated with Sources 1 and 2 were not consistently confirmed in the SIBTEST analyses. This could be attributed either to improper assignment of the items to the four sources or to the incorrect prediction of which language group an item would favour. The study, according to Gierl and Khaliq, was limited in that two mathematics DIF items and 10 social studies DIF items could not be associated with any of the four sources of translation DIF. Also, five of the 13 bundles in social studies were not correctly predicted. To clarify the source(s) of DIF in these items, further research was suggested in the area of the cognitive process underlying student responses using student interviews and protocol analysis (Ericsson & Simon, 1993).

Ercikan (2002), in a second study, attempted to disentangle sources of DIF in multilingual assessments using different strategies. Her second study focused on the comparability of English and French forms of the TIMSS mathematics and science using data from Canada (in English and French), England, France, and the United States. The study examined student performance on English and French versions of the test items in the Canadian administration and then cross-validated the findings in two other comparisons (i.e., England-France and United States-France) where the same versions of items were administered. The sample sizes ranged from 2,925 to 10,945 across the countries. DIF was detected using the IRT-based Linn-Harnisch (L-H) procedure (Linn & Harnisch, 1981) implemented using the PARDUX computer program (CTB/McGraw-Hill, 1991). The identification of adaptation and curricular differences as possible sources of DIF was attempted in three ways: a) judgmental review by bilingual translators, b) cross-validation of DIF in multiple groups, and c) investigation of the distribution of DIF by topic. Based on the average ratings given by the four translators in the judgmental review, 6 of the 22 mathematics DIF items (27.3%) for the Canadian English- and French-speaking examinees were interpreted to be associated with adaptation-related differences. This interpretation was also supported by the cross-validation analyses (in at least two out of the three comparisons). For science, 19 of the 52 DIF items (36.5%) were identified to be due to adaptation-related differences for the Canadian English- and French-speaking examinees. The existence of adaptation-related differences between the two language forms was further validated by the fact that 14 out of these 19 items (73.7%) were replicated in at least two of the comparisons. In addition, the judgmental review revealed three problems in adaptation leading to DIF: inadequate translation of key words, differential frequencies of the vocabulary used in the two languages, and discrepancies in the look and formatting of the item. Distribution of DIF items by topic provided curricular differences as a source of DIF for a small percentage of the DIF items--22.7% of the mathematics DIF items and 13.5% of the science DIF items. That is to say, half of the mathematics DIF items and half of the science DIF items were left unexplained by either adaptation-related or curricular differences. Ercikan (2002) concluded that judgmental review focusing on one or two sources should not be expected to explain the source of all of the DIF items, and therefore, that multiple sources need to be considered in examining sources of DIF, such as cultural differences and instructional differences.

In a third study, Ercikan, Gierl, McCreith, Puhan, and Koh (2004) examined the degree of comparability and sources of incomparability of the English and French versions of reading, mathematics, and science tests as part of the School Achievement Indicators Program (SAIP) in Canada. Data from the 1997, 1998, and 1999 administrations for 13- and 16-year-old students were employed. Two DIF detection methods were used to identify DIF: IRT-based Linn-Harnisch (LH) procedure and SIBTEST. Following statistical analyses of DIF, three strategies were used to identify adaptation and curricular differences as sources of DIF: (a) judgmental reviews by bilingual translators of all items, (b) cross-validation of DIF across two age groups, and (c) examination of distribution of DIF by curricular topic area. Without knowledge of the DIF status of the items, four bilingual translators were asked to independently rate the items in terms of the degree of equivalence between the two language versions. Consensus was then sought in the group discussions.

DIF analyses suggested that the L-H method consistently identified more DIF items than SIBTEST, while SIBTEST identified much larger numbers of Level-C DIF and more DIF items favouring English-speaking examinees. For the reading test, 18.2%

and 31.8% of the items were identified with DIF by both DIF detection procedures for 13- and 16-year-olds, respectively. All these DIF items were identified to be related to adaptation differences by the judgmental review. In mathematics, where only the L-H procedure was used, 37.6% and 32.0% of the items exhibited DIF for the two age groups, respectively. Among these DIF items, 36.2% and 37.5% were attributed to adaptation differences, respectively. In science, 36.1% and 34.0% of the items were identified as DIF by both methods for the two age groups, respectively. Among these DIF items, 53.8% and 44.9% were interpreted to have adaptation-related differences. In total, judgmental review associated 36.2% to 100% of the DIF items with adaptation-related differences in the three subjects across two age groups. The DIF items identified to be related to adaptation differences were replicated in both age group comparisons in 75.0% of the cases in reading, 52.9% of the cases in mathematics, and 60.7% of the cases in science. The replication of these DIF items in both age groups thus provided additional evidence supporting the interpretation that DIF might be due to adaptation differences. On the other hand, distribution of DIF by topic identified curricular differences as sources of DIF for 17% and 25% of the DIF items in mathematics for 13- and 16-year-olds, respectively, and 27% and 33%, respectively, of the DIF items in science. Ercikan et al. noted that a large proportion of DIF items could not be attributed to either adaptation-related or curricular differences, which corresponds to the findings reported by other researchers (e.g., Ercikan, 2002; Gierl et al., 1999; Gierl & Khaliq, 2001).

Identifying Sources of DIF in Adapted Tests

In the studies discussed above, a variety of DIF detection methods with diverse populations were used to examine the extent to which items function differentially between different language versions of a test after controlling for ability. Attempts were also made to identify the causes of DIF using substantive evidence. As documented by Ercikan (2002), the success rates for identifying sources of DIF in adapted tests varied. Ercikan (2002), for example, reported success rates of 27.3% and 36.5% on TIMSS mathematics and science, respectively. Ercikan (1998) found that 44.4% of the DIF items in an international assessment were linked to adaptation-related differences. Allalouf et al. (1999), on the other hand, reported much more encouraging results: their reviewers identified sources of DIF for more than 80% of the DIF items. Ercikan et al. (2004), more recently, reported a range of 36.2% to 100% DIF items across age groups and content areas on the SAIP tests that were attributed to adaptation-related differences.

Some common sources of DIF in adapted tests can be summarized from the studies discussed above (Allalouf et al., 1999; Ercikan, 1998; Ercikan, 2002; Gierl & Khaliq, 2001). They include: (a) differences in difficulty of vocabulary or sentences, such as differential frequency, difficulty, context meaning of vocabulary, omissions or additions that affect meaning, or differential length or complexity of sentences; (b) differences in content; (c) differences in format; (d) differences in cultural relevance. Among the four sources of DIF, the first source is most often associated with DIF identified using statistical procedures (Ercikan, 1998; Ercikan, 2002; Gierl & Khaliq, 2001)

As noted by some of these researchers (Ercikan, 2002; Ercikan et al., 2004; Gierl et al., 1999; Gierl & Khaliq, 2001), a large proportion of DIF items could not be attributed to either adaptation or curricular differences. For one thing, multiple factors other than adaptation-related or curricular differences should be considered in the study of sources of DIF in adapted tests. For another, there has been evidence to suggest that reviewers are not able to identify all the flaws in test items (Hambleton, 2005). Therefore, as pointed out by Gierl and Khaliq (2001), further research is needed in the area of the cognitive processes underlying the responses of the students, the intended population for the tests. The use of student interviews, as illustrated below, could help further understand the sources of DIF in adapted tests.

The Use of Think-aloud Interviews for DIF Analysis

Think-aloud protocols are structured interview protocols that require examinees to think aloud and talk about their understanding of test questions, the solution strategies they used, and the difficulties they encountered while answering test questions. After reviewing over 50 studies, Ericsson and Simon (1993) concluded that verbal reports collected under certain conditions can provide valuable and trustworthy information about cognitive processing during task performance. A number of researchers (e.g., Gierl, 1997; Hamilton, Nussbaum, & Snow, 1997; Katz, Friedman, Bennett, & Berger, 1996;

Leighton, Rogers, & Maguire, 1999) have used verbal reports to investigate the cognitive processes underlying students' responses to test items. Leighton (2005) made a strong argument for the collection of verbal reports in educational achievement testing by discussing the value of cognitive models for educational testing and addressing the misconception and misuse of verbal reports in the field. Hamilton et al. (1997) noted in particular how verbal reports provided valuable insights into students' cognitive processes that were not evident to researchers from simply reading the items.

A few studies have employed think-aloud protocols to help explain the presence of DIF. Hamilton (1999) combined an exploratory DIF study with a set of interview data to provide evidence concerning sources of gender differences on constructed-response (CR) items. The investigation focused on gender differences on Grade 12 CR science items administered as part of the National Education Longitudinal Study of 1988 (NELS:88). Logistic discriminant function analysis (Miller & Spray, 1993) was used to detect DIF on the CR items. A total of 25 high school students were asked to think aloud as they individually completed the four CR items and a subset of 16 multiple-choice items. After answering each item, they responded to a set of interview questions that elicited additional information concerning solution strategies and sources of knowledge. The protocols provided support for the hypothesis that eclipses (one of the CR items) has some dependence on visual or spatial reasoning, which tends to favour male students. The importance of knowledge acquired outside of school was also demonstrated, particularly for items that favoured male students. In other cases, although the DIF study did not provide clear guidance pertaining to which items should be considered biased, it did reveal that simple rules regarding content or format were insufficient to explain gender differences on science achievement tests. Above all, the study demonstrated the benefits of supplementing statistical analysis with an investigation of the cognitive processes that items elicit. With only one test examined, however, the generalizability of the findings to other CR items is limited.

Ercikan, Law, Arim, Domene, Lacroix, and Gagnon (2004) used think-aloud protocols for DIF analysis to investigate the English and French versions of Canada's national survey of achievement tests, School Achievement Indicators Program (SAIP) mathematics and science tests. The subjects were Grade 7 students from Vancouver, including 36 English-speaking and 12 French-speaking students. The think-aloud protocol consisted of a set of questions that tapped: (a) the participants' understanding of what the intent of each item was, (b) the steps that participants took to answer the item, (c) the reasons for selecting the answer that they chose, and (d) what aspects of the item facilitated or hindered the problem-solving process. Upon completion of each item, test administrators would ask only those questions that had not been spontaneously answered by the participants' think aloud process. A total of 20 mathematics and science items were selected for think-aloud. These items were statistically identified to function differentially between the two language groups, and judgmental review indicated that the sources of DIF were related to adaptation differences. The protocol data were used to determine whether the hypothesized sources of DIF were supported or whether other sources of DIF could be identified. The results provided supporting evidence for language differences as sources of DIF for seven out of the 20 items, six of which were the hypothesized sources of DIF. In particular, six types of responses were found to have provided support for hypotheses: a) students' reported understanding of questions; b) students' reading/misreading of test questions; c) students' reported familiarity/unfamiliarity with certain vocabulary or terminology; d) use of cues or miscues in their responses; e) students' different success levels on the test questions; and f) students' wrong responses. The authors concluded that think-aloud protocol analysis is a promising approach to disentangle sources of DIF, not as a preferred method but more as a complementary method to other methods such as judgmental review and statistical methods. Think-aloud protocols provide some unique evidence in support of the hypothesized sources of DIF that could not have been obtained using either judgmental review or statistical analysis.

The above two studies illustrate the use of think-aloud protocols in identifying the sources of DIF. They both demonstrate that protocol analysis can provide unique evidence in identifying the causes of DIF, and therefore should be used as a complementary tool in addition to statistical DIF analysis. The consistency between judgmental review and protocol analysis, however, was not high. In Ercikan et al. (2004), for example, the think-aloud protocols provided supporting evidence for language differences as sources of DIF for only seven out of the 20 items (35%), six of which were the hypothesized sources of DIF. Ercikan et al. (2004) noted that the discrepancy might be due to failure in eliciting the kind of responses from students that would support the hypotheses established in judgmental reviews, or to the limitations of the sample of students used in the study.

In the present study, the French versions of the tests were developed for Grade 9 French-speaking students in Alberta, 90% of whom were from French Immersion programs (Jolanta Wojcik, personal communication, June 8, 2005). A brief overview of French Immersion education, therefore, is presented next to provide some contextual information. Emphasis is placed on the case in Alberta, especially in regard to French Immersion students' language proficiency as well as the effect of language of testing on their performance.

Overview of French Immersion Education in Canada

French immersion programs were introduced into Canadian schools in the 1970s to encourage bilingualism across the country. French Immersion is a school program in which students who have little of no prior contact with French are put together in a classroom setting in which French is used as the medium of instruction. It involves teaching subjects partly or entirely in French to students whose mother tongue is English.

French Immersion is administered in many different fashions in every province and territory. The two predominant forms are Early Immersion and Late Immersion. Early Immersion, a program started in Kindergarten or Grade One, is the most common delivery model in Alberta. The students learn 100% of their curriculum for the first two or three years fully in French. English Language Arts are typically introduced in Grade 3. The French load diminishes as students progress to higher grades. In Grades 7 to 9, for example, about 50% to 80% of the curriculum is delivered in French. From Grades 10 to 12, about 40% to 80% of the class time is in French. Late Immersion is a program that begins around the commencement of Junior High School (e.g., Grade 7). For optimum language development, from 90% to 100% of class time is spent in French in the first few months. In the months and years to follow, recommended instructional time in French is the same for Late Immersion as for the Early Immersion at the same grade levels. Late Immersion is not as intensive as Early Immersion, and deemed not as effective (Alberta Learning, 2002a).

The 2000 Program for International Student Assessment (PISA) results provided data on the extent of enrollment in French Immersion programs among 15-year-old students in each of the ten Canadian provinces. Enrollment ranged from 2% in British Columbia to 32% in New Brunswick, and girls accounted for 60% or more of students in all provinces except Quebec. In Alberta, 4% of 15-year-olds were enrolled in French Immersion programs, and 80% of them were from Early Immersion programs.

In 1992, the Canadian Education Association summarized the effectiveness of the French Immersion programs in this way:

> French immersion is a proven successful Canadian approach to second language learning.... No educational program has been so intensively researched and evaluated in Canada as has French immersion. The effects of the program on the acquisition of French-language as well as English-language skills, and the academic achievement of French immersion students, have been well documented and research shows that the program works. (p. 2)

Language Proficiency

In Alberta, the desired outcomes for French Immersion students in regard to language proficiency are "a high proficiency in the English Language" and "functional fluency in French" (Alberta Learning, 2002a, p. 1). In other words, their level of proficiency in English language is expected to be equivalent to English program students. In French, however, they do not reach native-like competence by the end of Grade 12 in spite of their ability to communicate effectively (Alberta Learning, 2002b). This is the usual result in French Immersion programs in Canada (Genesee, 1987).

English language proficiency. Research on the academic achievement of French Immersion students has shown that after an initial lag lasting until a year or two after English Language Arts is introduced, Early French Immersion students perform as well in English as their English-program counterparts (Edwards, 1989). There is evidence that from late elementary on, early immersion students may outperform their English-program counterparts in some English skills. On Grade 3, 6, and 9 English language Arts achievement tests, for example, students in French Immersion programs tend to perform better than non-French Immersion students (Alberta Learning, 2002a).

Data from PISA (2000) showed that in every province, except Manitoba, students enrolled in French Immersion programs outperformed their counterparts in non-immersion programs in reading performance when tested in English (Allen, 2004). While it is true that students in French immersion are generally from higher socio-economic backgrounds, when gender, socio-economic background, and parents' education were each taken into account, French Immersion students still outperformed their counterparts in regular English programs (Allen, 2004). Therefore, it seems safe to conclude that English reading achievement is higher among French Immersion than non-immersion students.

French language proficiency. By the end of Grade 12, French Immersion students generally achieve a high level of functional fluency in French, which enables them to participate easily in conversations in French, pursue post-secondary education with French as the language of instruction, and accept employment where French is the language of work (Alberta Learning, 2002a). However, when compared with students whose first language is French, the skills of French Immersion students are below those of Francophones, especially in speaking and writing (Holobow, Genesee, Lambert, & Chartrand, 1987; Swain & Lapkin, 1981).

It is not hard to understand why French Immersion students fail to reach native-like proficiency in French considering the fact that French Immersion students typically have little exposure to French outside of the classroom. Research has shown that French Immersion students are more likely to read, watch television, and communicate with peers and adults in English rather than in French when they are out of school (Cummins, 1987; Swain & Lapkin, 1981). More recently, Romney, Romney, and Menziers (1995) examined how much reading 127 Grade 5 French Immersion students in Alberta did for pleasure in both French and English. They found that more than two-thirds of the French Immersion students never read for pleasure in French outside school. The average amount of time they devoted to reading in French was 25 minutes a week as compared to 183 minutes a week in reading in English. Likewise, they watched considerably less television in French (8 minutes per week) than in English (478 minutes per week). The study also suggested that after almost six years of schooling in French, the vast majority (85%) of the students surveyed claimed that they found reading in English easier, and the main difficulty for them to read in French was vocabulary.

Similarly, in an earlier study, Romney, Romney, and Braun (1989) found that French Immersion students' knowledge of words related to out-of-school activities was limited and this impeded their reading considerably. Difficulties with vocabulary were also most frequently given as the reason for French Immersion students' inferior performance in science achievement tests in British Columbia when compared with students in the regular English programs (Day & Shapson, 1996). Above all, when second language learning is limited to school experiences, students rarely achieve a native-like command of that language (Carey, 1981; Swain, 1974).

Effect of Language of Testing

In Alberta, Grades 3, 6, and 9 students have written provincial exams in mathematics for many years. In addition, Grades 6 and 9 students have also written provincial tests in science and social studies. In all the three subject areas, French Immersion students regularly show levels of achievement that are higher than the provincial levels for tests written in English (Alberta Learning, 2002b). In spite of their superior performance, French Immersions students' disadvantage in writing French tests as compared to English tests in language intensive areas has been demonstrated in earlier research. As illustrated in the study by Morrison and Pawley (1983), while their sample of Grade 9 French Immersion students did equally well in mathematics whether they were tested in English or French, Grade 10 French Immersions students performed less well in history when they were tested in French than when they were tested in English. The researchers suggested that the results could be attributed to knowledge of technical vocabulary and reading comprehension difficulties.

Similarly, Samuel (1990) investigated the effect that language of testing had upon the scores of French Immersion students writing a standardized test of Grade 6 social studies achievement test. The French and English forms of the test were randomly assigned to 179 French Immersion students: 95 wrote the English version and 84 wrote the French version. The results revealed that French Immersion students achieved significantly lower scores when they wrote an achievement test in French as compared to English. In particular, the effect sizes on topic specific data-based questions were all larger than the effect sizes on the same topic discrete item reporting categories.

In addition, Alberta Education (1990, 1991, & 1992) undertook a three-year study to determine if language of testing was a variable that affected the way French Immersion students responded to test questions. A group of Grades 3 and 6 students (ranging from 232 to 693) were randomly assigned to write English or French versions of the achievement tests in mathematics, science, and social studies. The results suggested that student responses at both grade levels were sensitive to the language of testing. In all three subject areas, French Immersion students who wrote the English forms of the achievement tests achieved significantly higher scores than French Immersion students who wrote the French forms of the tests (see Table 1). The difference between the two groups was generally greater at the Grade 3 level than it was at the Grade 6 level. For the Grade 6 level, in particular, the difference was greatest in social studies ($M_E = 69.9$ vs. $M_F = 53.8$), less in science ($M_E = 42.9$ vs. $M_F = 36.9$), and least in mathematics ($M_E = 39.3$ vs. $M_F = 37.3$). In addition, 517 Grade 9 students took the social studies achievement test (259 in English and 258 in French). Their responses also revealed that

Table 1

Language of Testing Study (1990-1992) Results

	Total	Mean		Standard Deviation		t^a	p^b
	Possible	English	French	English	French		
Grade 3							
Mathematics	50	36.2	32.3	7.8	8.7	6.25	.000
Science	50	33.9	27.6	8.8	9.2	7.52	.000
Social Studies	50	29.3	21.8	8.4	8.0	-	-
Grade 6							
Mathematics	55	39.3	37.3	8.9	9.9	2.86	.004
Science	60	42.9	36.9	8.7	9.8	7.53	.000
Social Studies	100	69.9	53.8	14.8	14.5	-	-
Grade 9							
Social Studies[c]	60	42.3	38.6	8.7	9.7	4.6	.000

Note. [a] *t* test statistic. [b] Probability level. [c] Multiple-choice items only.

French Immersion students who wrote the English form of the test achieved significantly higher scores than French Immersion students who wrote the French form ($M_E = 42.3$ vs. $M_F = 38.6$). The difference, however, was considerably smaller than the differences at lower grades. It was concluded that the trend of smaller differences with the progress of grades might be due to either a better command of French by Grade 9 or the drop-out of students with a weak command of French before Grade 9, or both.

To recapitulate, French Immersion students are able to achieve a high level of functional fluency in French by the end of Grade 12, but they are not likely to acquire native-like command of French. Writing tests in French, therefore, might depress their performance on the tests.

Literature Summary

Five main conclusions can be drawn from the literature discussed in this chapter. First, in spite of their wide use, forward translation and back translation have drawbacks attributable to the fact that a source language version is established before a target language version. Second, the simultaneous test development approach is promising in creating more equivalent tests in that it tends to maximize linguistic and cultural decentering and minimize the chances of obtaining nonequivalent items. Third, differential item functioning is an effective tool in evaluating the equivalence of adapted test items. Fourth, the success rate for identifying sources of DIF on adapted tests using judgmental review varies, and the sources of DIF may be related to adaptation, curriculum, or reasons that cannot be identified. Fifth, protocol analysis can provide unique evidence in identifying the causes of DIF and, therefore, be used as a complementary tool in addition to statistical DIF analysis and judgmental review.

Consequently, the present study is justified in employing a combination of judgmental review, statistical analysis, and think-aloud interviews in an effort to investigate the equivalence of achievement tests in English and French developed using the simultaneous approach. It is hypothesized that adaptation contributes marginally to the occurrence of DIF, and the simultaneous test development approach is efficacious in reducing differential item functioning attributable to adaptation differences and enhancing linguistic and cultural decentering.

CHAPTER THREE

OVERVIEW OF STAGE 1: ITEM DEVELOPMENT AND PILOT TESTING[1]

As mentioned earlier, this doctoral dissertation is part of a large-scale research project designed to investigate the validity and utility of the simultaneous approach to the development of equivalent achievement tests in French and English. The research project was initiated in 2002, and the first stage of the study was reported in Rogers, et al (2003). To provide some necessary background information for the present study, the first stage of the project is reviewed in this chapter.

Item Development

Subject Areas and Grade Level

Grade 9 mathematics and social studies were selected for the purpose of this research. Gierl et al. (1999) and Gierl and Khaliq (2001) found that adaptation differences were more pronounced in social studies, a language-rich content area, than in mathematics. The social studies curriculum is more sensitive to differences in cultural values and preferences than the mathematics curriculum. It was hoped that by including both subjects, the findings in one content area would help illuminate the findings in the other content area. Grade 9 was selected as the grade level because evidence suggested that Grade 9 students were quite capable of verbalizing their thoughts and justifying their answers to test questions (Leighton et al., 1999). This skill is critical for the think-aloud procedures that were completed during the fourth stage.

Item Writers

Six item writers were recruited to develop the items for mathematics and social studies, three for each subject. They were all from the greater area of Edmonton in Alberta and nominated by officials in the Learning Assessment Branch of Alberta Education. As shown in Table 2, there were two male teachers and one female teacher on each subject writing team. French was the first language for one (Item writer A) of the item writers for mathematics and all the three writers for social studies. One mathematics

---

[1] This chapter was adapted from Rogers et al. (2003).

Table 2

Background of Item Writers

	Mathematics			Social Studies		
Characteristic / Item Writer:	A	B	C	D	E	F
Gender	FE	MA	MA	FE	MA	MA
First language	F	E	E	F	F	F
Language used daily	F&E	E	E	F	F&E	F&E
Years of teaching experience	23	7	7	15	15	13
Years of teaching mathematics/social studies	7	5	5	8	9	1
Language used to teach	F	F	F	F	F&E	F
Language competence[a]						
in French	5	4	4	5	5	5
in English	5	5	5	4	3	3
Knowledge and understanding of[a]						
Curriculum	4	5	4	5	4	4
Instructional procedures	5	4	4	5	4	4
Knowledge of[a]						
Culture specifics of French	5	4	3	5	5	5
Culture specifics of English	5	5	5	4	3	3
Cross-culture psychology	4	4	3	4	4	4
Test development background						
Completed an educational assessment course	No	No	No	Yes	Yes	Yes
Item writer for provincial testing program	Yes	No	No	Yes	No	No
Language used	E			F		
Previous translation experience	No	No	No	No	No	No
Knowledge of test development[a]	3	4	3	4	4	3

Note. FE = Female, MA = Male; F = French, E = English. [a] Self-ratings of knowledge (1 = very weak, …. , 5 = very strong). From "Differential Validity and Utility of Successive and Simultaneous Approaches to the Development of Equivalent Achievement Test in French ad English," by Rogers et al., 2003, *The Alberta Journal of Educational Research*, *49*, p. 299.

item writer (A) used both French and English daily; the remaining two used English. Two social studies item writers (D and F) used both languages daily while the third used French. All but one (F) had taught the subject for which they developed items for at least five years. They were all teaching Grade 9 French Immersion classes at the time of item writing.

While they were all confident about their French-language competence, two writers for social studies (E and F) were not sure about their competence in English. The six item writers were all confident about their knowledge of the curriculum and the instructional procedures to follow. Whereas three writers (A, B, and D) were confident about their knowledge of shared meanings and cultural specifics of French and English and cross-cultural psychology, item writers C, E, and F were unsure about cultural specifics in their second language.

When it comes to their background in test development, the three item writers for mathematics had not completed an educational assessment course, while the three members for social studies had. One writer on each team (A and D) had experience writing items for the provincial achievement testing program. None had previous translation experience. Last, three item writers (B, D, and E) were confident about their level of knowledge about test development, whereas the other three were less sure.

Item Writing

The definition of the constructs for the tests was based on the "level of thinking-by-subject matter" table of specifications for Alberta Grade 9 achievement tests in mathematics and social studies (see Appendices A and B). A short version of the *Taxonomy of Educational Objectives: Cognitive Domain* (Bloom, 1984) was reviewed in order for the item writers to better understand the level-of-thinking dimension in the table of specifications. To facilitate item development, a set of guidelines for constructing multiple-choice items (Hopkins, Stanley, & Hopkins, 1990) were reviewed and discussed, along with the four common types of translation errors identified on previous provincial achievement tests (Gierl & Khaliq, 2001). Finally, the nature of the item-writing task was explained. The item writers were asked to write one item at a time. They could choose to write the item first in either English or French. The writer immediately translated the item into the second language. They were not allowed to move on to the next item until they had made sure that the items in both languages meant the same and called for the same level of thinking by English- and French-speaking students.

All of the above preparation activities were completed in half a day. Then, each item writer independently developed 30 items in two and one half days. Each day started at 9:00 am and ended at 4:00pm, with a one-hour lunch break and coffee breaks determined by the item writers. The item writers on each team continuously interacted during the item construction phase. Their discussions focused mostly on the ways to translate words and expressions in one language to the other.

Reactions to Simultaneous Test Development

At the end of the third item-writing day, the item writers were asked to provide their feedback about the simultaneous test development process they engaged in. First, they were presented with the following statement:

> Some people claim that one big advantage of the simultaneous approach is that it ensures maximum linguistic and cultural comparability in the definition of the construct and the test items designed to measure it.

They were then asked to indicate the degree (1 = strongly disagree, ..., 5 = strongly agree) to which they agreed with this statement with respect to linguistic and cultural comparability. The results were somewhat mixed. Two of the three members on each team either agreed or strongly agreed with the above statement with respect to linguistic comparability. The third mathematics item writer (B) was not sure while the third social studies item writer (E) disagreed. When it comes to cultural comparability, the three mathematics item writers indicated they were unsure. In contrast, two of the social studies item writers agreed that the simultaneous development approach led to cultural comparability while the third item writer (E) was not sure. The discrepancy between the two teams with respect to cultural comparability may be attributable to differences between the nature of mathematics and social studies. The mathematics item writers were not sure how the French and English cultures were differentially involved. In contrast, culture and the values within culture form an important part of social studies.

Frequency of changes. The item writers were asked about the frequency with which they changed the item as first written when writing it in the second language. It should be noted that while the item writers were allowed to write the items in either language first, all but one item were first drafted in French. The item writers pointed out that since they were teaching either mathematics or social studies in French at the time, it was only natural to do so.

In general, the mathematics teachers made changes less frequently than did the social studies teachers. This result is not surprising given the fixed nature of mathematics compared to social studies. All item writers appreciated the opportunity to make changes during the first item development stage. Two reasons were provided. First, the item writers commented that any weaknesses in an item showed up immediately instead of later in the translation process. Second, the item writers indicated that the item meaning was retained in both languages due to the immediacy of the translation or, as one of the teachers put it, "the essence and objectives [to which questions are referenced] are fresh in our minds." The discussions that took place during the item writing revolved around the meaning of a word in one language and the comparability of the meaning of the corresponding word in the other language.

Difficulty of simultaneous development. The item writers were then asked to indicate how difficult they found the task of simultaneously developing items in both French and English. A five-point Likert scale (1 = not difficult at all, …, 5 = very difficult) was used for this purpose. The three ratings for the mathematics teachers were 1, 2, and 2; and the three ratings for the social studies teachers were 2, 4, and 3. That is to say, mathematics item writers found the task of simultaneous test development relatively easy, while social studies item writers found it somewhat difficult. As one social studies teacher commented: "Translation in English was quite challenging at times, and brought me back at times to modify the French version." This difference between the two subject areas can be attributed to the fact that social studies items tend to involve more vocabulary than mathematics items.

Strengths and weaknesses of simultaneous test development. The item writers on both teams indicated the following attributes as strengths of the simultaneous test development approach:

1. efficiency and speed;
2. reduced loss of comparable meaning because one version was written immediately after the other;
3. better assurance that the level of language in both forms is suitable and incidental vocabulary does not confuse the students;
4. immediacy of the process;

5. helps us to be as specific as we can be in both languages;
6. done at the same time by the same person, thereby avoiding differences that come up when one person prepares an item in one language and a second person does the translation; and
7. allowed for continual revision of each item.

The item writers identified four weaknesses, two of which were applicable to all test adaptation methods and two of which were specific to the simultaneous development approach. The first two were the tendency to translate literally to the detriment of linguistic integrity and the challenge to translate published quotations and tabular information from one language to another. The two concerns particular to the simultaneous approach included the need to keep in mind and maintain a sharp focus across both cultures and the need for teachers who are really familiar with the curriculum in both languages.

Item Review and Revision

Following the item development, the two subject area teams met separately with two research team members two weeks later to review and discuss each item. One of the research team members was fluently bilingual and possessed strong knowledge of the shared meanings and cultural specifics of the French and English languages and culture and cross-culture psychology. The second research team member possessed expertise in the area of measurement and evaluation. Attention was paid to the comparability between the two language versions of each item in terms of meaning, and the correctness and the appropriateness of writing in each language version. The nature of simultaneous test development was retained throughout the revision process: one item was addressed at a time and both versions were equally open to modification.

The review and revision for mathematics took approximately four hours. The same task for social studies took nine hours in total, split into two sessions. The changes made included correcting grammatical errors in one or both languages, changing awkward expressions in one language but not the other, and word translation errors. Due to the less dependence on language in mathematics as compared with social studies and the greater objectivity of mathematics than of social studies, the mathematics teachers made changes much less frequently than social studies teachers. In the revision process, the simultaneous approach was preserved. Both language versions were equally open to modification, and efforts were made to avoid word-for-word translation and awkwardness in language.

Placement of Items in the Table of Specifications

Following the review and revision process, the placement of the items in their respective table of specifications was reviewed. Of particular concern was the placement of the items according to the level of thinking required. Several mathematics items that assessed similar thinking levels were placed at both thinking levels within the table of specifications for mathematics. This was not the case for social studies.

Consequently, the mathematics item writers met to review the placement of their items along the level of thinking dimension. Altogether, they made 25 changes. Five changes involved moving an item to a different topic (e.g., from numbers to patterns and relations). The remaining changes involved level of thinking classification: three items were reclassified at the higher level and 17 were reclassified at the lower level. The discussion and reassignments centered on mathematical procedures and whether they were known and could be applied "automatically" or whether some conscious thought was required. If it was the former, the item was classified at the knowledge level; otherwise it was classified at the skill level (see Appendix A).

Pilot Testing

The purpose of the pilot test was to determine the item characteristics to be used to guide further revision. The intent was not to test the equivalency of the forms at this point given that the pilot tests was conducted in March 2003 and not toward the end of the school year when all the coursework would have been completed.

Pilot Test Forms

A total of 87 mathematics items and 86 social studies items survived the review and revision process. Given that the pilot testing was to occur in one class period (50 minutes), the items were divided into two forms for each subject. After grouping the items by thinking level in each topic area, the initial draft of the mathematics pilot test forms contained 35 items, and the initial draft of the social studies forms contained 39 items.

The two teams of item writers met again to examine the equivalence of expressions and meaning of items for each form of the pilot tests. The changes made included correcting the spelling and accents in French for both mathematics and social studies. As a result of this review, four items were deleted from one social studies form and five were deleted from the other form. Among them, four were deleted because of lack of clarity in both languages, three were deleted because of lack of match between tabled source information and the questions, and two were deleted because of the lack of a clear reproduction of what was initially a colored map. Lastly, the item writers examined the items in the pool not included in the pilot test forms and were asked if any of these items should replace an item in the pilot tests. No changes were made. The final numbers of items in the mathematics forms were 35, and social studies forms contained 34 and 35 items, respectively.

Given the date of the pilot tests, the teachers in the sample classes would not have covered all the material in the curriculum (see Appendices A and B). Additionally, although all teachers in the province must teach the same material, not all teachers follow the same sequence when teaching the subject area topics. Consequently the students in the different classes would be exposed to different topics. Therefore, the teachers of the sampled classes completed a form on which they indicated whether they had taught, were presently teaching, or still needed to teach each of the subject area topics.

Pilot Test Samples

The pilot test forms were administered in the French Immersion classes of the six item writers. To control for school effects, an English-speaking class in their schools was also administered the English version of the test forms. The forms were counter-balanced to control for any class effects.

Pilot Test Results

The sample sizes for the pilot test forms ranged from 26 to 53 (see Table 3). In spite of the small sample sizes, the item analysis completed using LERTAP (Nelson, 2000) provided valuable information for the next round of revision along with the feedback from the teachers in whose class the tests were administered. As shown in Table 4, the items were classified into three classes according to their item characteristics. Class A included items with the discrimination index (the uncorrected point-biserial) of at least 0.20 for both language groups. Class B contained 1) items for which the discrimination index was at least 0.20 for one language group and most of the teachers for the other language group indicated that the topic had not been taught or was being taught at the time, and 2) items for which the discrimination index was less than 0.20, but positive for both groups, and the topic had not been taught or being taught at the time. Class C contained the remaining items. Generally, Class A items were considered good, Class B contained fair items, and Class C items were considered problematic. The distributions of the items by class across the topic areas for each subject area are presented in Table 4.

Table 3

Pilot Test Sample Sizes

	Content Area			
	Mathematics[a]		Social Studies	
Form	French	English	French	English
1	26	36	43	50
2	28	38	44	53

Note. [a] Although the teachers were asked to tell the students to answer all questions and to do their bests, the mathematics teachers in one school advised their students either to answer the items they wished or to answer only the questions that were related to material they had been taught. The data for the students of these teachers was incomplete. Consequently the responses from this school were not included in the analysis. From "Differential Validity and Utility of Successive and Simultaneous Approaches to the Development of Equivalent Achievement Test in French ad English," by Rogers et al., 2003, *The Alberta Journal of Educational Research*, *49*, p. 300.

Table 4

Distribution of Items by Class

Mathematics	Topic								
	Number		Patterns & Relations		Shapes & Space		Statistics & Probability		
Level of Thinking	K	S	K	S	K	S	K	S	Total
Item Class									
A	2	8	2	6	3	2	1	3	27
B	3	1	2	8	3	9	1	3	30
C	0	2	0	4	1	3	3	0	13

Social Studies	Topic								
	Technology & Change		Economic Systems		Quality of Life		Former USSR		
Level of Thinking	K	S	K	S	K	S	K	S	Total
Item Class									
A	3	6	7	9	0	3	2	1	31
B	3	1	6	5	3	6	0	2	26
C	1	0	2	3	0	4	1	1	12

Note. From "Differential Validity and Utility of Successive and Simultaneous Approaches to the Development of Equivalent Achievement Test in French ad English," by Rogers et al., 2003, *The Alberta Journal of Educational Research, 49, p.* 301.

Mathematics. Of the 70 mathematics items, 27 items were in Class A, 30 items were in Class B, and 13 items were in Class C. Inspection of the distributions of item difficulties for Class A and Class B items within each language group revealed that the distributions were essentially uniform. The corresponding means and standard deviations were 0.49 and 0.22 for French and 0.44 and 0.17 for English, Class A and 0.46 and 0.20 for French and 0.32 and 0.16 for English, Class B. The observations that the item means for both groups are lower than those typically found on the provincial tests is attributable to the time of year the pilot tests were administered (March and not June). There was no significant difference between the item means for French-speaking and English-speaking students in Class A. For Class B, however, the item means for the French-speaking students were higher than the corresponding means for the English-speaking students ($p < 0.01$). This finding may be attributable to the fact that the French-speaking students were French Immersion students and that, as reported by the teachers, these students tend to have high socioeconomic status.

The sample sizes were not large enough to control for ability and conduct differential item functioning analysis. Rather, the intent of the pilot study was to obtain preliminary information on the performance of the items. This information revealed that, given the number of items in Class A and Class B, the range of difficulty for both language groups, and the distribution of the items across the cells of the table of specifications, it would be possible to construct a mathematics examination of 40 relevant and representative items which, when administered toward the end of the year, would yield means and standard deviations commensurate with end-of-year performance.

Social studies. Thirty-one of the 69 social studies items were in Class A, 26 items were in Class B, and 12 items were in Class C. As for mathematics, the distributions of item difficulties for Class A and Class B items within each language group were essentially uniform. The corresponding means and standard deviations were 0.68 and 0.16 for French and 0.49 and 0.13 for English, Class A and 0.48 and 0.24 for French and 0.32 and 0.15 for English, Class B. As for mathematics, the item means for both groups were lower than those typically found on the provincial tests due to the time of the year at which the pilot test was conducted. Further, for both Class A and Class B, the item means for the French-speaking students were higher than the corresponding means for the English-speaking students ($p < 0.01$). This finding again appears to be attributable to the fact that the French Immersion students tend to have high socioeconomic status. However, the intent of the pilot study was to obtain preliminary information on the performance of the items. This information again revealed that it would be possible to develop a social studies examination of 40 relevant and representative items which, when administered toward the end of the year, would yield means and standard deviations commensurate with end-of-year performance.

Discussion

The evidence collected through the item development stage suggested that the simultaneous test development method allowed the influence and integration of information from item writers and reviewers representing different language and cultural groups to affect test development directly. The discussions that took place extended beyond the simple choice of comparable words and phrases to the form of expressions in each language and whether differences in form would be allowed in an attempt to maintain comparable meaning while recognizing the idiomatic differences between the two languages. Both the French and English versions of each test were equally open to modifications. Evidence suggested that item writers were able to give deeper consideration to subtle language and culture issues in the item development process.

What was not clear was why the mean performance of the French-speaking students was significantly better than the mean performance of the English-speaking students. One hypothesis was that the differential performance was attributed to socioeconomic differences between the two language groups. A second hypothesis was that the two versions of each item were not as comparable as initially thought. Consequently, the purpose of this dissertation was to examine the tenability of these two hypotheses. In particular, emphasis was placed on the investigation of the degree to which differential item functioning (DIF) existed between the two language groups and the degree to which DIF could be attributed to adaptation-related differences.

CHAPTER FOUR

METHODS AND RESULTS: JUDGMENTAL REVIEW AND ITEM SELECTION

As indicated earlier, the present study is part of a large research project. The main purpose of this research project was to investigate the validity and utility of the simultaneous approach to the development of equivalent achievement tests in French and English. The first stage of the research project, which involved test development, item revision and pilot testing, was reported in Rogers et al. (2003) and was reviewed in the previous chapter. The present study comprised the second, third, and fourth stages, focusing on the compilation of evidence to evaluate the comparability of the Grade 9 mathematics and social studies tests in English and French developed using the simultaneous approach.

During the second stage, the item writers met to revise the items based on the item analysis of the pilot test results. Then a panel of six certified translators reviewed the retained items from the revision for comparability in meaning and wording. Revisions were again made based on the comments of the translators. Lastly, one mathematics test (28 items) and one social studies test (40 items) were assembled in both languages.

In the third stage, the four test forms were administered to a sample of Grade 9 students as part of the field-testing conducted by Alberta Education. The student responses on the field tests were scored and item analyses were performed using the LERTAP item analysis computer program (Nelson, 2000). The factor structures of the tests were then examined using NOHARM (Fraser, 1988) and LISREL 8.14 (Jõreskog & Sörbom, 1996). Differential item functioning (DIF) analyses were conducted using SIBTEST (Shealy & Stout, 1993).

The fourth stage involved explaining the DIF found in Stage 3. First, teachers of the classes included in the field test samples were asked to comment on each item with respect to item clarity, relation to the learner outcomes, and curriculum coverage. The teacher comments were examined to determine whether the identified DIF was attributable to adaptation differences or to other sources. Second, a sample of DIF items and non-DIF items were used for think-aloud interviews. Protocol analysis (Ericsson &

Simon, 1993) was conducted to compare the patterns in which students from the two language groups understood the items.

Given the sequential nature, with each step used in the second, third, and fourth stages dependent on the results of the previous step, the procedures used and the results for each stage are presented together. That is, the methods and results for Stages 2, 3, and 4 are presented in Chapters 4, 5, and 6, respectively.

Item Revision by Item Writers

Based on the item analysis results derived in the first stage, each team of item writers and members of the research team met to revise the items in the pilot tests (70 items for mathematics and 69 for social studies). Again, the two language forms were equally open to changes: the results for each item were considered in both languages before proceeding to the next item. Items that could not be revised to satisfaction were deleted.

Of the 70 items in the mathematics tests, 12 items were deleted. Altogether, 21 English items and 27 French items were revised. Among them, 18 items were modified in both versions. Of the 18 common items, three items were revised because the correct answer was not included in the options by mistake; five items were modified to make the options more plausible and, thus, more appealing; nine items were revised to make the items more concise and clear; and one item was modified in order to make the two versions more equivalent in meaning. The remaining three English items were changed to match the meaning or wording in the French version, and the nine remaining French items were revised to match the meaning or wording in the English version. In general, the revisions made were minor. For example, in one English item, the phrase *Mary Ann is observing a ship from a cliff that is 120m high* was changed to *Mary Ann is observing a ship from the top of a 120m cliff.* This change was made to better reflect the French version: *Marianne observe un navire du haut d'une falaise de 120 m.* In the French version of another item, *La circonférence de la Terre à l'équateur mesure* $4{,}0 \times 10^4$ *km*, *mesure* [measure] was changed to *est* [is]. This change was made to better match the English version: *The circumference of the earth at the equator is* 4.0×10^4 *km.* As a result, 58 mathematics items were retained.

Among the 69 social studies items, 18 were deleted. Altogether, 28 English items and 30 French items were revised. Of these items, 24 items were modified in both versions to make the items more concise and clearer. For example, one question was originally phrased as *What are the two important factors that contributed the most to the industrialization process in England?* The research team felt that the wording was somewhat awkward and unclear, so the question was changed to *What are the two most important factors that contributed to the industrialization process in England?* Corresponding changes were made in the French version. Similarly, to emphasize the period of time as demonstrated in a table, the question *Which American state was the least successful in attracting Canadian immigrants during this period* was changed to *Which American state was the least successful in attracting Canadian immigrants from 1997 to 2000.* The French version was also revised to reflect this modification.

In addition to the 24 items that were revised in both versions, four English items were revised to match the meaning or wording of the French version, while six French items were modified to match the English version. In general, the revisions made to these items was not substantial. For example, in one English item, *according to this paragraph* was changed to *according to this text.* This change was made to better match the French version, *selon ce texte.* In the French version of another item, *Un système économique qui met la* ***plus grande*** *emphase sur la recherche du profit est le système d'économie..., la recherche du profit* [in search of profit] was changed to *le profit* [profit]. This change was made to better match the simplicity of wording in the English version, *An economic system that has the* ***greatest*** *emphasis on profits is the....* Following these revisions, 51 items were retained in social studies.

Item Review by Certified Translators

Six accredited translators were recruited to review the items retained from last round of revision. They were members of the Association of Translators and Interpreters of Alberta (ATIA), which is an association affiliated with the Canadian Translators and Interpreters Council (CTIC) and the International Federation of Translators. To retain their status as a certified translator, the members must pass the national CTIC examination once every three years.

The six translators were all female. They independently evaluated each of the 58 mathematics items and 51 social studies items in terms of the comparability in meaning of the two language forms. A three-point scale (1 - different, 2 - similar, and 3 - identical) was used to assess the comparability of meaning. Five of the six translators also participated in the evaluation of the comparability of wording of each item. Specifically, a *Yes* or *No* format was used to assess the comparability in words, phrases, verb tenses, or form of expression (i.e., idiom). For each item that was identified to be different in meaning or wording, the translator was asked to justify her ratings in the questionnaire and provide suggestions to correct the problem. These comments and suggestions formed the basis for the subsequent revision and deletion of items.

Inter-rater Agreement on Comparability of Meaning

The degree of agreement among the translators for the comparability of meaning was examined using two statistics: the judges' discrepancies from the median (*JDM*; Rogers, 2001) and item ambiguity (Rogers, 2001). While the *JDM* provides a measure of a rater's discrepancy from the median of all raters across all the items, item ambiguity quantifies rater agreement on each item. The formula for *JDM* is as follows:

$$JDM_j = \sum_{k=1}^{K} \left| X_{kj} - Md_k \right|$$

where X_{kj} is the rating given by judge j to item k, Md_k is the median of the ratings given by the J judges to item k, and K is the number of items. If there is perfect agreement among the raters, each rater's *JDM* will be zero. Raters with a *JDM* exceeding the *JDM*s for the remaining raters by a considerable amount are considered aberrant raters and might, therefore, be removed from subsequent analysis. Next, the inclusive range, *R*, of the ratings for each item (Rogers, 2001) provides the measure of item ambiguity. *R* equals the difference between the highest and lowest ratings for the item, plus one (i.e., the inclusive range):

$$R = H - L + 1.$$

A value of 1 indicates that the highest and lowest ratings are the same; values of 2 or 3 indicate some disagreement among the raters.

The *JDM* values for the six translators are presented in Appendix C for mathematics and Appendix D for social studies. Inspection of the values revealed that the *JDM*s for mathematics ranged from 4.5 to 17.5, while the *JDM*s for social studies ranged from 15.5 to 33.5. Rater 5 had the highest *JDM* in both subjects, 6 points and 9 points above the second highest *JDM*. This indicated that her ratings were very different from the ratings of other translators. During a meeting with the principal researcher of the research project, she pointed out that there was a format difference throughout the mathematics tests: English numbers had a decimal period while French numbers had a decimal comma. This finding together with other comments she made raised concerns that she did not understand the language differences between English and French and what the adaptation process entailed. Therefore, Rater 5 was removed from all subsequent analyses.

The *JDM* and *R* values for the remaining five raters are presented in Table 5 for mathematics and Table 6 for social studies for each of the remaining raters. The new *JDMs* ranged from 5 to 11 for mathematics and 11 to 24 for social studies, indicating better agreement in mathematics than social studies. In term of *R*, 38 items (65.5%) obtained a 1 in mathematics, indicating perfect agreement. Twelve items (20.7%) obtained a *R* of 2, indicating moderate agreement, while eight items (13.8%) obtain a *R* of 3, indicating poor agreement. For five of the eight items with poor agreement, however, the large values were due to one rating of 1 for that item. Therefore, the overall inter-rater agreement in mathematics was very good. For social studies, eight items (15.7%) obtained a *R* of 1, indicating perfect agreement. Twenty-eight items (54.9%) obtained a *R* of 2, indicating moderate agreement, while 15 items (29.4%) obtained a *R* of 3, indicating poor agreement. Similar to mathematics, for 9 of the 15 items with poor agreement, the large values were due to one rating of 1 for that item. On the whole, considering the content nature of social studies, it was not surprising that the translators did not agree as frequently for the social studies items as they did for the mathematics items in regard to comparability of meaning. Given this, the overall agreement in social studies was deemed to be satisfactory.

Table 5

Mathematics Translator Review Results

Item	Rater 1	Rater 2	Rater 3	Rater 4	Rater 6	Median	Range
1	3	3	3	3	3	3	1
2	3	3	3	3	3	3	1
3	1	3	3	3	3	3	3
4	3	3	3	3	3	3	1
5	3	1	3	3	3	3	3
6	3	3	2	3	3	3	2
7	3	3	3	3	3	3	1
8	3	3	3	3	3	3	1
9	3	3	3	3	3	3	1
10	3	3	3	3	3	3	1
11	3	3	3	3	3	3	1
12	3	2	3	3	3	3	2
13	3	3	3	3	3	3	1
14	2	2	3	3	3	3	2
15	3	3	3	3	3	3	1
16	3	3	3	3	3	3	1
17	3	1	2	2	2	2	3
18	1	1	3	1	3	1	3
19	3	3	3	3	3	3	1
20	3	3	3	3	3	3	1
21	3	3	3	3	2	3	2
22	3	3	3	3	3	3	1
23	3	3	3	3	3	3	1
24	3	3	3	3	3	3	1
25	3	3	3	3	3	3	1
26	3	3	3	3	3	3	1
27	3	3	3	3	2	3	2
28	3	3	3	1	1	3	3
29	1	1	1	1	1	1	1
30	3	3	3	3	3	3	1
31	3	3	3	3	3	3	1
32	3	1	3	3	3	3	3
33	3	3	3	3	3	3	1
34	3	3	3	3	3	3	1
35	3	3	3	3	3	3	1
36	3	3	3	3	3	3	1
37	3	3	3	3	3	3	1
38	1	1	1	1	1	1	1
39	3	3	3	3	3	3	1
40	3	3	3	3	3	3	1
41	3	3	3	1	3	3	3
42	3	2	3	3	3	3	2
43	3	2	3	2	3	3	2
44	3	3	3	3	3	3	1
45	3	3	3	3	3	3	1
46	3	3	3	3	3	3	1
47	3	3	3	3	3	3	1
48	3	3	3	3	2	3	2
49	3	3	3	3	3	3	1
50	3	3	3	3	3	3	1
51	3	1	3	1	1	1	3
52	3	3	3	3	2	3	2
53	3	3	3	3	3	3	1
54	3	2	3	3	2	3	2
55	3	3	3	3	2	3	2
56	3	3	3	3	3	3	1
57	3	3	3	3	3	3	1
58	3	2	2	3	3	3	2
JDM	6	11	6	5	10		

Note. Code for comparability of meaning: 1 = different, 2 = similar, and 3 = identical. Range = highest rating – lowest rating + 1.

Table 6

Social Studies Translator Review Results

Item	Rater					Median	Range
	1	2	3	4	6		
1	3	2	2	3	3	3	2
2	3	2	3	3	3	3	2
3	2	1	2	1	2	2	2
4	2	2	3	2	3	2	2
5	3	3	3	3	3	3	1
6	3	3	2	3	2	3	2
7	2	2	2	2	2	2	1
8	3	3	3	3	1	3	3
9	3	1	3	1	2	2	3
10	3	2	2	3	3	3	2
11	3	2	3	3	2	3	2
12	3	3	3	3	3	3	1
13	3	2	3	3	3	3	2
14	3	2	3	3	3	3	2
15	3	3	3	3	2	3	2
16	3	3	3	3	3	3	1
17	3	1	3	3	3	3	3
18	1	3	3	3	3	3	3
19	3	2	3	3	2	3	2
20	2	1	2	3	3	2	3
21	3	2	3	3	3	3	2
22	3	2	3	3	3	3	2
23	3	1	1	3	2	2	3
24	3	3	3	3	2	3	2
25	1	1	1	2	2	1	2
26	3	3	3	3	1	3	3
27	3	2	3	3	3	3	2
28	3	3	1	3	2	3	3
29	2	3	2	3	3	3	2
30	3	1	1	2	3	2	3
31	2	3	3	2	3	3	2
32	3	1	2	2	2	2	3
33	3	3	2	3	2	3	2
34	2	2	2	2	2	2	1
35	3	2	3	3	3	3	2
36	2	3	3	3	3	3	2
37	2	3	3	3	1	3	3
38	3	3	3	3	3	3	1
39	2	2	2	2	3	2	2
40	2	3	3	3	3	3	2
41	1	2	3	2	3	2	3
42	3	3	3	3	-	3	1
43	1	1	2	2	2	2	2
44	1	1	1	3	3	1	3
45	2	1	1	3	3	2	3
46	3	3	3	3	3	3	1
47	3	2	3	2	3	3	2
48	2	2	2	3	3	2	2
49	3	1	3	1	1	1	3
50	3	3	2	3	2	3	2
51	3	2	2	2	2	2	2
JDM	16	22	16	11	24		

Note. Code for comparability of meaning: 1 = different, 2 = similar, and 3 = identical. Range = highest rating – lowest rating + 1.

Comparability of Meaning

The degree to which the raters as a panel felt the meanings of the French and English versions were the same was assessed using the median of the judges' ratings. A median equal to 3 indicated that the five judges rated the two language forms as identical while a value of 1 indicated the five judges rated the two language forms as different. The median was selected rather than mean because of the small number of raters and because the median is not sensitive to outliers. The median value for each item is presented in Table 5 (mathematics) and Table 6 (social studies). A summary is provided in Table 7.

Table 7

Comparability of Meaning for Mathematics and Social Studies Items

	Mathematics		Social Studies	
Median	No. of Items	Percentage	No. of Items	Percentage
1	4	6.9	3	5.9
2	1	1.7	15	29.4
3	53	91.4	33	64.7

Mathematics. As shown in Table 7, the median value was 3 for 53 (91.4%) mathematics items. Item 17 obtained a median of 2 because the English version used a short form of *algebra-tiles* (alge tiles) which the translators deemed non-existent in English. The median rating for the four remaining (Items 18, 29, 38, and 51) mathematics items was 1, meaning that these items were not equivalent in meaning. A review of the translators' comments pointed to typographical errors in three of the four items (the four options for Item 18 were -43, -7, - 3, -7 in English, and -43, -37, -13, -7 in French; the name *Joe* was mistakenly replaced with *Sue* in French for item 29; The option C for Item 51 was *SSA* in English, and *ACC* in French, which was equivalent to *ASS* but not *SSA* in English). For the fourth item, Item 38, the first part of the English version said *Consider the model shown below*, while the French version said *Pour le modèle ci-dessous identifie l'élévation gauche* [For the model below, identify the left elevator]. Thus, extra information, *identify the left elevator*, was given in French. Combined with other differences in wording (e.g., missing word in French), this item was deemed different in meaning by all five translators. Further, the research team agreed that the diagrams used in this item might be confusing to some students. As a result, Item 38 was deleted.

Social studies. The median value was 3 for 33 of the social studies items (64.7%) and 2 for 15 items (29.4%) (see Table 7). These 15 items were interpreted as similar in meaning with minor differences. For example, the English version of Item 30 said *who expresses concern*, while the French version said *Qui manifeste **le plus** d'inquiétude* [who expresses the most concern]. Item 43 said *was it most likely* in English and *est-il plausible* [is it possible] in French: these two expressions did not have the same meaning for three of the reviewers. The median rating for the remaining 3 items (Items 25, 44, and 49) was 1. A review of the translators' comments pointed to one typographical error in one of the three items: the word *exigerait* was mistakenly typed as *sigerait* in the French version of Item 25. Item 44 was deemed different due to differences in wording in the reading material, such as *closing the hospital beds* in English vs. *la fermeture des hôpitaux* [closing the hospital]. For Item 49, four translators identified that option B in the two language versions did not agree on the meaning. In English, it said *the government's rejection of socialism created uncertainty for some Russian citizens*, while in French it said *le rejet du communisme représente une incertitude pour des citoyens russes* [the rejection of communism represents uncertainty for some Russian citizens].

Comparability of Wording

Five translators independently evaluated all the items in mathematics and social studies for comparability of wording, one of whom was Rater 5. Therefore, only the comments made by four translators were used for subsequent review and revision. As expected, very few language differences were found in mathematics, while more differences were identified in social studies. For mathematics, out of the 928 possible cells (58 items × 4 wording categories × 4 raters), 54 cells (5.8%) were flagged showing language differences. For social studies, of the 816 cells (51 items × 4 wording categories × 4 raters), 137 cells (16.8%) were flagged.

To sum up, mathematics items were more comparable in meaning than social studies items. Although the proportions of items that were considered comparable in meaning (identical or similar) were about the same for mathematics (93.1%) and social studies (94.1%), mathematics contained considerably higher percentage of items that were considered identical in meaning than social studies (91.4% vs. 64.7%). In terms of wording, not surprisingly, more differences existed in social studies than mathematics (16.8% vs. 5.8%). Given the language-rich nature of social studies tests, however, item comparability was deemed to be satisfactory.

Item Revision and Selection for Field-testing

Item Revision

Following item review, all the language differences identified by the four translators were considered, and changes were made to the items where deemed necessary. Three members of the research team completed these revisions. The nature of simultaneous test development was maintained throughout the revision process: an item revised in one language was then checked in the other language to ensure comparability before moving to the next item. Two items in mathematics (Items 38 and 51) and one item in social studies (Item 23) that differed in meaning were deleted due to the difficulty in achieving equivalency between the English and French versions.

The remaining revisions were not extensive. For mathematics, the typographical errors identified earlier for Items 18, 29 and 51 were corrected. Other changes were minor such as ensuring two numbers to the right of the decimal point and leaving space between the amount of money and the dollar sign in French. More items were revised in social studies, but the changes were minor. For example, Item 6 in English originally asked *What are the two most important factors that contributed to the industrialization process in England?* The word *process* was deleted because the French equivalent did not appear in the French version and the removal of the word did not change the meaning of the English version. In Item 1, all the options in English contained the word *their*, but the French equivalent was not included in the French version. To be concise, *their* was deleted from all the options in the English version.

English	*French*
1. What was the impact of the implementation of mass production on the workers?	**1.** Quel effet a eu l'implantation de la production en série sur les travailleurs?
The specialization of work and	La spécialisation du travail et
A. a reduction in their working hours.	A. la réduction des heures de travail.
B. their improved working conditions.	B. l'amélioration des conditions de travail.
C. their gain of more control over production.	C. plus de contrôle sur la production.
D. a loss of their control over the end product.	D. la perte de contrôle sur le produit final.

Similarly, Item 13 showed a difference in verb tense, *is* versus *serait* [would be]. To match the English version, *serait* was changed to *est* [is] in French.

English	*French*
13. In this scenario, Speaker II is in favour of a society essentially based on a	**13.** Dans ce scénario, l'interlocuteur II serait en faveur d'une société essentiellement basée sur
A. Free market economy.	A. une économie de marché libré.
B. mixed market economy.	B. une économie mixte.
C. Centrally planned economy.	C. une économie planifiée.
D. traditional economy.	D. une économie traditionnelle.

Item Selection

Following the item revision, 56 mathematics items and 50 social studies remained. Out of these items one test consisting of 28 mathematics items and one test containing 40 social studies items were constructed in both English and French. These numbers matched the numbers in the pilot tests developed by Alberta Education in these two subject areas. Item selection was based on the test specifications for each subject area, with the items distributed proportionally as close as possible in terms of content areas (topics) and thinking levels (knowledge or skills). Items not in need of revision were selected first in each cell of the tables of specifications followed by the revised items. For mathematics, the task of selecting 28 items was relatively easy given the pool of 56 items. For social studies, however, 40 items needed to be selected from the pool of 50. Considering that some items were too similar in nature and/or format and therefore could not been selected at the same time, the room for selection was not large. The final test forms for both subjects were believed to be comparable in meaning and wording across the two language versions. Copies of mathematics test forms are included in Appendices E and F and copies of social studies test forms are included in Appendices G and H.

CHAPTER FIVE

METHODS AND RESULTS: FIELD TESTING

This chapter describes the statistical procedures performed on the field test data of the English and French versions of Grade 9 mathematics and social studies tests. The procedures and results for descriptive analysis, dimensionality assessment, and differential item functioning are presented respectively.

Field-test Subjects

In Alberta, English-speaking students represent the dominant language and cultural group. Students who receive instruction in French represent only 5.0% of the Grade 9 population, including students in French Immersion programs (4.5%) and Francophone students (0.5%) (J. Wojcik, personal communication, June 8, 2005). French Immersion programs are typically operated within English-speaking schools, but the language of instruction is French in many subject areas. For Grades 7 to 9, for example, typically 50% to 80% of the curriculum is delivered in French. The immersion programs are designed for students whose first language is not French but who want to become functionally fluent in French. In contrast, Francophone students attend French schools where the language of instruction is all French. French schools are designed for students with at least one French-speaking parent. Students in French schools are expected to master French as a mother tongue, and establish a sense of identity and belonging to the French community. Francophone students write the provincial achievement tests in French. Although teachers of French Immersion students can opt for either French or English as the language of testing for their students, very often French is chosen because the language of instruction is French. Based on the Alberta Grade 9 provincial achievement test reports (Alberta Education, 2004), 2076 French Immersion students took the social studies test in French, while only 43 French Immersion students opted to take the test in English.

The final forms of the mathematics and social studies tests were administered to samples of Grade 9 students stratified by region as part of the field-testing conducted by the Learner Assessment Branch, Alberta Education in May 2004. For mathematics, 469 answer sheets (from 19 schools) were returned for the English form and 345 (from 12 schools) were returned for the French form. For social studies, 470 answer sheets (from 19 schools) were returned for the English form and 263 (from 12 schools) were returned for the French form.

Descriptive Analysis

Prior to beginning the analysis it was noted that the Learner Assessment Branch altered the wording of one reading text in social studies and two items in mathematics in the French forms prior to field-testing. In the case of social studies, the changes were in the text material that preceded the questions related to that material. While the changes made were not correct, it was felt that these changes would not alter the equivalence of the meaning between the English and French versions of the item. However, this was not the case for mathematics: English words were mistakenly used instead of French words in the French version of Items 7 and 10 (see Appendix F). The word "par" in French was replaced by "by" in English for the options in Item 7 and "et" was replaced by "and" in the options for Item 10. Therefore, these two items (Items 7 and 10) were deleted from all subsequent analyses.

The student responses were scored and analyzed using the LERTAP computer program (Nelson, 2000). LERTAP is an item and test analysis program based on classical test theory. Classical test score analysis was chosen for this study because of the relaxed requirement regarding sample sizes. Item Response Theory (IRT) models require at least 500 examinees, depending on the number of item parameters included in the model (Lord, 1980).

As shown in Table 8, the psychometric characteristics of the tests and items of the two language versions were generally comparable except for the mean scores. First, the mean test scores for the French-speaking examinees are significantly higher than the means for the English-speaking examinees in both mathematics ($t(812) = 5.81,\ p < 0.01$) and social studies ($t(731) = 4.25, p < 0.01$). The corresponding effect sizes (Glass & Hopkins, 1995, p. 290), which were obtained by dividing the mean difference of the two samples by the standard deviation of the English-speaking samples (i.e., the English group was considered the control group), were of moderate size: $d = 0.41$ for

Table 8

Psychometric Characteristics of the Mathematics and Social Studies Tests

	Mathematics		Social Studies	
	English	French	English	French
No. of Examinees	469	345	470	263
No. of Items	26	26	40	40
Mean	13.14	15.04	23.74	25.75
Standard Deviation	4.64	4.56	6.11	6.12
Skewness	0.36	0.02	-0.14	-0.34
Kurtosis	-0.38	-0.47	-0. 61	-0.67
Internal Consistency [a]	0.76	0.75	0.79	0.81
Item Difficulty: Mean	0.50	0.57	0.59	0.64
Item Difficulty: SD	0.15	0.16	0.17	0.17
Item Discrimination[b]: Mean	0.28	0.28	0.34	0.28
Item Discrimination: SD	0.09	0.09	0.11	0.14

Note : [a] Cronbach's alpha. [b] Point-biserial Correlation.

mathematics and *d* = 0.33 for social studies. Second, the standard deviation, kurtosis and skewness values were similar between the two language groups for each test, indicating that the test score distribution were comparable across language groups. Third, the internal consistencies (Cronbach's alpha) were comparable across groups, 0.76 vs. 0.75 for mathematics and 0.79 vs. 0.81 for social studies. When the number of items is increased to the number of items typically included in the corresponding provincial test (50 items for mathematics and 55 items for social studies), the internal consistencies are equivalent to 0.85 for both subjects, based on the Spearman-Brown Prophecy Formula (Brown, 1910; Spearman, 1910). Fourth, corresponding to the discrepancies in mean total score, the mean item difficulty for the French group was also higher than that for the English group, in both mathematics and social studies. Lastly, mean item discrimination was calculated by transforming the point-biserial correlations for each item using Fisher's *z*-transformation, summing the transformed correlations, dividing by the total number of items, and calculating the antilog to convert back to the mean point-biserial correlation. In mathematics, the two language versions were comparable in regard to mean item discrimination. In social studies, however, the English version was somewhat more discriminating than the French version. On the whole, the psychometric characteristics of the English and French versions of the tests were comparable, with the one exception being that the French sample outperformed the English sample in both mathematics and social studies.

Structural Equivalence Analysis

Before conducting the Differential Item Functioning (DIF) analysis to assess differences in performance at the item level between the English- and French-speaking students after controlling for ability, the validity of the matching criterion (in this case, the total test score) must be defended by ruling out construct bias (Sireci, 1997; Zumbo, 2003). In the psychometric literature, factor analysis is commonly used to evaluate the construct equivalence across groups at the test level. If construct equivalence is established, DIF analysis can be used to examine differential performance at the item level (e.g., Ercikan & Koh, 2005; Gierl, Rogers, & Klinger, 1999; Reise, Widaman, & Pugh, 1993; Sireci, Fitzgerald, & Xing, 1998). Therefore, a combination of factor analysis and DIF analysis was used in this study to examine the comparability of two test forms in mathematics and social studies.

Non-linear Factor Analysis

To determine the factor structures of the four test forms, nonlinear factor analysis was used (McDonald, 1967)[2]. The nonlinear factor analysis was conducted using the program NOHARM (Fraser, 1988). NOHARM allows an exploratory approach and a confirmatory approach. If the underlying dimensional structure is not known, then the exploratory mode of NOHARM should be used. If a particular dimensional structure is hypothesized with firm supportive evidence, then the confirmatory mode should be used. In the present study, the exploratory approach was taken, beginning with one factor and then seeing if the addition of factors led to a better solution as assessed by Tanaka's

---

[2] Linear factor analyses such as principal component and principal axis extraction with phi coefficients have been found to indicate more factors than are actually present in the data (Hambleton & Rovinelli, 1986; Nandakumar, 1994). The use of tetrachoric correlations can lead to non-positive definite matrices.

(1993) unweighted least squares goodness-of-fit index and the root mean square residual (RMSR). Tanaka's index takes a value of 1.0 if the model under consideration fits the data perfectly and 0.0 if the fit is no better than chance. There are no interpretative guidelines for Tanaka's index, except that a higher value implies better model fit. The RMSR has no upper bound but takes smaller values as fit improves and has a value of zero if the model is a perfect fit to the data. A RMSR equal to or less than four times the reciprocal of the square root of the sample size implies good model fit (Fraser, 1988). This RMSR criterion ranged from 0.18 to 0.25 in the present study. The most conservative RMSR value, 0.18, was adopted.

The fit indices for the 1- and 2- dimensional models for the mathematics and social studies test forms are presented in Table 9. For both language versions of the mathematics and social studies tests, the unidimensional model fitted the data well: the changes in the two fit statistics were marginal when the number of factors was increased from one to two. In particular, Tanaka values went up by 0.007 to 0.013, and RMSR values went down by 0.000 to 0.001.

Table 9

NOHARM Fit Indices for 1- and 2-Dimensional Models

No. of Factors	Mathematics				Social Studies			
	English		French		English		French	
	Tanaka	RMSR	Tanaka	RMSR	Tanaka	RMSR	Tanaka	RMSR
1	0.964	0.010	0.951	0.012	0.946	0.010	0.920	0.012
2	0.973	0.009	0.960	0.011	0.953	0.010	0.933	0.011

Multi-group Confirmatory Factor Analysis

In the psychometric literature, confirmatory factor analysis (CFA) is a widely used approach for examining whether the factor structures of a test are invariant across two or more language groups (Lietz & Roche, 1996). In the present study, multi-group CFA was employed to evaluate further whether the factor structure of the test data was consistent across English and French versions. Following the procedures employed by Gierl (2000), Gierl et al. (1999), Meara and Sireci (2003), and Sireci et al. (1998), parcels of two or more items were created to serve as units of analysis in the CFA. Two main reasons underlie the use of parceling in the present study. First, the use of parcel scores better meets the normal-distribution assumption underlying the use of maximum likelihood estimation in CFA, especially when the item-level responses are dichotomous. It has been recognized that item data can be combined to optimize the normality of data (Cattell, 1956; Cattell & Burdsal, 1975; Gorsuch, 1983). Second, due to differences in item difficulty and examinee item responding strategies, the error associated with a single multiple-choice item is generally large (Dorans & Lawrence, 1987). Parceling can produce stronger indicators with higher reliability (Cattell, 1956; Cattell & Burdsal, 1975; Dorans & Lawrence, 1987, Gierl, 2000).

In the present study, parcels were created by summing items in each curricular content area by cognitive level cell in the test specifications. The test specifications guided test development and characterized the test developers' representations of the content areas and cognitive skills measured by the test forms. The items developed for each cell in the test specifications are similar in content and cognitive coverage, and therefore are relatively homogenous. In this study, mathematics had four content areas and two cognitive levels. However, only seven parcels could be constructed since for one content area (statistics and probability) all four items measured skills. Similarly, while social studies had four content areas and two cognitive levels, seven parcels were again constructed because one content area (the former USSR) had only one item at the knowledge level. In this case, this item was combined with other items in the content area to form one parcel. These parcels served as unit of analysis in the CFA of structural equivalence.

The LISREL 8.14 program (Jöreskog & Sörbom, 1996) was used to complete the multi-group CFA. A full measurement invariance model was tested by equating the number of factors, factor loadings, and error variances associated with the factor loadings across the two language groups for mathematics and for social studies. Only the one-factor model was assessed given the results of the non-linear factor analysis (see Table 9). A number of goodness-of-fit indices are currently available to assess confirmatory factor analytic models. However, there is little agreement on which index provides the best answer to the question of model fit (Bollen & Long, 1993; McDonald & Marsh, 1990). Therefore, multiple indices were used to assess the model fit. The first index was the chi-square statistic, which indicates whether the restrictive hypothesis tested can be rejected. A model is considered to have acceptable fit if the difference between the variance-covariance matrix generated by the original data and by the hypothesized model is small, yielding a nonsignificant chi-square (e.g., $p < 0.05$). Although the chi-square test is sensitive to large sample sizes (Bentler & Bonett, 1980), the chi-square statistic is one of the most frequently used fit indices in structural analyses for educational research (Elliott, 1994; Gierl & Mulvenon, 1995). The second index used was the root mean square error of approximation (*RMSEA*), which indicates the "badness" of the model per degree of freedom. That is, the *RMSEA* is a measure of fit that adjusts for parsimony by assessing the discrepancy per degree of freedom in the model. The third index used was the root mean square residual (*RMR*), which measures the average size of the residuals when the model is fitted to the data. For both the *RMSEA* and *RMR*, values of 0.05 or less indicate close fit of a model, and values of 0.08 reflect reasonable fit of a model (Reise et al., 1993). The fourth index used was the goodness of fit index (*GFI*), which is a measure of the amount of variance and covariance in the data accounted for by the model. Lastly, the adjusted goodness of fit statistic (*AGFI*) was used, which is a variant of *GFI* that adjusts for the degrees of freedom of the model by replacing the total sum of squares with the mean squares. For both the *GFI* and *AGFI*, the common lower bound is 0.90 for a good fit (Meara & Sireci, 2003).

The results of the multi-group CFA are presented in Table 10. Inspection of the fit indices indicated that the one-factor model fitted the data well for both in mathematics and social studies, even when the number of factors, factor loadings, and error variances were set equal across the two groups. Neither chi-square statistic is significant at the 0.05 level of significance; the *RMSEA* and *RMR* indices are all below 0.05; and the *GFI* and *AGFI* are well above 0 .90.

Taken together, the results of the non-linear factor analysis at the item level and the CFA at the parcel level revealed that each language version of both the mathematics and social studies tests was unidimensional with comparable factor loadings and error variances. Consequently, it was possible to conduct the DIF analysis.

Table 10

Tests for Model Equivalence between English-speaking and French Immersion Examinees

Content Area	χ^2	*df*	*RMSEA*	*RMR*	*GFI*	*AGFI*
Mathematics	44.47	41	0.014	0.043	0.98	0.98
Social Studies	49.02	41	0.023	0.047	0.97	0.97

Note. RMSEA = root mean square error of approximation; RMR = root mean square residual; GFI = goodness of fit index; AGFI = adjusted goodness of fit index.

Differential Item Functioning (DIF)

To evaluate whether and to what extent differences existed between the performance of the two groups of students on each item, an exploratory three-step approach (Camilli & Shepard, 1994; also see Roussos & Stout, 1996a; Ramsey, 1993; Zieky, 1993) was used. First, the simultaneous item bias test (SIBTEST: Stout & Roussos, 1999) was used to identify items for which there were differences, if any, in performance between the two groups. Next, to identify the sources of DIF, each item that displayed DIF was examined employing data from the teacher comments and student think-aloud protocols. Last, an item was considered to be biased if it was established that the source of the unexpected or "extra" difficulty for one group was not relevant to what the test measures (e.g., adaptation differences); otherwise the source of the DIF was considered to be undeterminable and in need of further research (Camilli & Shepard, 1994).

The simultaneous item bias test (SIBTEST) is a nonparametric statistical method for assessing DIF of an item or a bundle of items. It is based on Shealy and Stout's (1993) multidimensional model for DIF. The basic assumption is that multidimensionality produces DIF. SIBTEST detects bias by comparing the responses of examinees in the reference and focal groups that have been allocated to the same bins using their scores on a "matching subtest" (Stout & Roussos, 1995). The matching subtest is a subset of items that, ideally, are known to be unbiased.

In SIBTEST, the amount of DIF in the studied item is reflected in the effect size estimate $\hat{\beta}_{uni}$, which is the weighted sum of the differences between the proportion-correct true scores on the studied item for examinees in the two groups across all score levels. The true scores are estimated using linear regression and then adjusted using a regression correction technique (Shealy & Stout, 1993). The weighted mean difference between the reference and focal groups on the studied item across the *k* subgroups is given by:

$$\hat{\beta}_{uni} = \sum_{k=0}^{k} p_k d_k ,$$

where p_k is the proportion of focal group examinees in subgroup *k* and d_k is the difference in the adjusted means on the studied item between the reference and focal groups, respectively, in each subgroup *k*. The statistical hypothesis tested by SIBTEST is:

$$H_0 : \beta_{uni} = 0 \qquad versus \qquad H_1 : \beta_{uni} \neq 0 .$$

The test statistic for evaluating the $\hat{\beta}_{uni}$ null hypothesis is:

$$SIB = \frac{\hat{\beta}_{uni}}{\hat{\sigma}(\hat{\beta}_{uni})},$$

where $\hat{\sigma}(\hat{\beta}_{uni})$ is the estimated standard error of $\hat{\beta}_{uni}$. SIB has a standard normal distribution with a mean of zero and standard deviation of one under the null hypothesis. The null hypothesis is rejected when $|SIB| > z_{1-\frac{\alpha}{2}}$. A statistically significant value of that is positive indicates DIF against the focal group and a negative value indicates DIF against the reference group. In the present study, the English sample was the reference group, and the French sample was the focal group.

Roussos and Stout (1996b) proposed general guidelines for interpreting the magnitude of item DIF: (a) negligible or A-level DIF: Null hypothesis is rejected and $\left|\hat{\beta}_{uni}\right| < 0.059$; (b) moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq \left|\hat{\beta}_{uni}\right| < 0.088$; and (c) large or C-level DIF: Null hypothesis is rejected and $\left|\hat{\beta}_{uni}\right| \geq 0.088$.

SIBTEST was selected for use in this study mainly for three reasons. First, SIBTEST uses a regression estimate of the true score as the matching variable, so that the examinees are matched on a latent rather than an observed score (Gierl et al., 1999). This has been shown to be useful in controlling the Type I error rate (Roussos & Stout, 1996a; Shealy & Stout, 1993). Second, the number of students assessed in the field test was between 263 to 470 across the four language/subject test forms, thereby precluding the use of parametric procedures like item response models that require larger numbers of students. Third, a number of studies have suggested that SIBTEST is more powerful in detecting DIF than other non-parametric procedures (e.g., Mantel-Haenszel) and parametric procedures (e.g., logistic regression) that are not dependent on large sample sizes (Bolt & Stout, 1996; Gierl, Jodoin, & Ackerman, 2000; Gierl et al., 1999; Jiang & Stout, 1998). For the purpose of this study, identification of more DIF items may result in a more thorough analysis of the test items, and thereby lead to a more comprehensive evaluation of the equivalence of the tests across languages.

SIBTEST Results

The SIBTEST results for mathematics and social studies are summarized in Table 11. Five out of the 26 (19.2%) mathematics items displayed DIF, while 17 of the 40 (42.5%) social studies items displayed DIF. Not surprisingly, social studies, a vocabulary-rich content area, contained a larger percentage of DIF items than mathematics.

Table 11

SIBTEST Results for the Mathematics and Social Studies Tests

	Mathematics			Social Studies		
	B-level	C-level	Total	B-level	C-level	Total
No. of DIF Items	2	3	5	8	9	17
No. of items favouring English	1	2	3	6	4	10
No. of items favouring French	1	1	2	2	5	7
Percentage of DIF items (%)	7.7	11.5	19.2	20.0	22.5	42.5

DIF: Mathematics

As shown in Table 12, two of the five DIF items in mathematics, were identified with moderate DIF ($0.059 \leq \left|\hat{\beta}_{uni}\right| < 0.088$) and three items were identified with large DIF ($\left|\hat{\beta}_{uni}\right| \geq 0.088$). Three DIF items (one moderate and two large) favored English-speaking examinees, and two DIF items (one moderate and one large) favored French-speaking examinees.

Table 12

Distribution of DIF in the Mathematics Tests

Item Number	$\hat{\beta}_{uni}$	DIF Level	Favouring Group
8	0.089	C	English
9	-0.142	C	French
11	0.112	C	English
12	0.077	B	English
24	-0.086	B	French

DIF: Social studies

As shown in Table 13, of the 17 items identified with DIF, 8 were of moderate size and 9 were of large size. Ten (6 moderate and 4 large) favoured English-speaking examinees, and seven (2 moderate and 5 large) favored French-speaking examinees.

Table 13

Distribution of DIF in the Social Studies Tests

Item Number	$\hat{\beta}_{uni}$	DIF Level	Favouring Group
1	0.069	B	English
3[a]	0.141	C	English
4[a]	0.123	C	English
9	0.064	B	English
11[a]	-0.130	C	French
12[a]	-0.109	C	French
13	-0.104	C	French
17[a]	-0.116	C	French
18[a]	-0.078	B	French
19	0.083	B	English
26[a]	0.060	B	English
28[a]	0.081	B	English
31	-0.082	B	French
36[a]	-0.107	C	French
37	0.132	C	English
38	0.088	C	English
39	0.084	B	English

Note. [a] Items that are chained (sharing the same stimulus text with one or more items)

CHAPTER SIX
METHODS AND RESULTS:
TEACHER COMMENTS AND THINK-ALOUD INTERVIEWS

This chapter presents the procedures and results for teacher comments and think-aloud interviews. First, teacher comments were collected during field-testing to see if any adaptation or curriculum differences could be identified between the English and French versions of the tests. To clarify the extent to which adaptation differences contributed to the DIF found in the last stage, think-aloud interviews were conducted with English-speaking and French Immersion students. The results from both procedures are also reported in this chapter.

Teacher Comments

Following the DIF analysis, the teacher comments collected during field-testing were analyzed in order to see whether the teachers identified any adaptation or curriculum differences. As part of the field-testing procedure used by Alberta Education, teachers whose classes completed the field tests were asked to comment on the test items, identifying problems they saw with the items and content areas they had not yet covered. For mathematics, 15 of the19 English teachers commented on the English form and 10 of 14 French Immersion teachers commented on the French form. For social studies, 18 of the 19 English teachers provided comments for the English form and 9 of 11 French Immersion teachers for the French form. Their comments are summarized in Table 14 for mathematics and Table 15 for social studies. For linguistic problems, only those suggestions that were deemed correct by the bilingual team member of the research project were included.

Mathematics

The teachers' comments on the mathematics tests did not provide plausible explanations as to why the five items displayed DIF. As shown in Table 14, Item 12 received the most comments. Ten of the English teachers who made comments about the English test thought this question was long and, therefore, hard to understand, while only

Table 14

Teacher Comments on the Mathematics Tests

Number	DIF Level	Group Favoured	English (15)	French (10)
4	-	-		*dollars canadienne → dollars canadiens (1)*
5	-	-	Do they have to break even or make a profit? (1)	Do they want to make a profit or just cover the $650? (1)
8	B	E	Not covered (1)	
9	C	F		*ci-dessous* should be *ci-dessus* (1)
12	B	E	Complicated wording(10) Not covered (1)	Complicated wording (1)
14	-	-	Not covered (1)	
20	-	-	Not covered (1)	
21	-	-	Not covered (1)	
23	-	-	Not covered (3)	
25	-	-	Not covered (5)	
26	-	-	Not covered (5)	
27	-	-	Not covered (1)	
28	-	-	Not covered (2)	

Note. E= English, F = French. Number in brackets indicates the number of teachers who made the comment.

one French Immersion teacher made similar comments about the French version. One English-speaking teacher indicated that the content tested in this item had not been covered in her class. No evidence was thus found explaining DIF in Item 12 favoring the English group.

Eight non-DIF items (14, 20, 21, 23, 25, 26, 27, and 28) were noted by at least one English-speaking teacher as "not covered", including all of the four items tapping the topic of Statistics and Probability (25, 26, 27, and 28). In spite of the difference in curriculum coverage, none of these items were differentially difficult for the two language groups. In regard to item clarity, one teacher from each language group questioned about the clarity of Item 5. It was not clear to them whether the students' union wanted to make a profit or just cover the cost.

Lastly, linguistic problems were identified in one DIF item and one non-DIF items. For Item 4, *dollars canadienne* should be changed to *dollars canadiens* to match *Canadian dollars* in English. For Item 9, one French-speaking teacher identified an error regarding the position of the diagrams included: *ci-dessous* [below] should be changed to *ci-dessus* [above]. This error was made by the Learner Assessment Branch that conducted the formatting and printing of the test. They moved the diagram up, and changed *below* to *above* in the English version, but forgot to make the corresponding change in the French version. In spite of this error, Item 9 displayed C-level DIF favouring the French group.

Social Studies

The teacher comments on social studies tests are summarized in Table 15. The comments provided plausible explanations for four of the 17 DIF items. For example, Items 3 and 4 both favoured English-speaking examinees. While the English-speaking teachers made no comments about the texts or questions, two French-speaking teachers found the structure of the third paragraph of the stimulus material in the French version confusing, which might have attributed to the DIF found in Items 3 and 4. For Item 17, which favoured French-speaking examinees, three English-speaking teachers commented that they had not yet covered the concept of interventionism, which happened to be the correct answer. No French-speaking teachers made similar comments. Therefore, DIF of Item 17 might be due to curriculum differences: at least 3 of the 20 English-speaking classes had not studied interventionism. For the fourth item, Item 39 that favoured English-speaking examinees, two typographical errors were identified in the correct answer, option B, which might have contributed to the DIF in this item.

In addition to Item 17 discussed earlier, two English-speaking teachers noted that universality referenced in Item 35 had not been discussed in class. Similarly, two French-speaking teachers said its counterpart, *universalité*, was a hard word for Grade 9 students. However, this item did not display DIF.

Table 15

Teacher Comments on the Social Studies Tests

Item Number	DIF Level	Group Favored	English (18)	French (9)
1	B	E	Implementation is a hard word (1) Wordy & confusing (3)	*implantation → la mise en œuvre* (1) Ambiguous question (2)
3	C	E		Structure of the 3rd paragraph is confusing (2)
4	C	E		
5	-	-		
6	-	-	Confusing (2)	Need revision (2)
11	C	F	*anarchy* is a hard word (1)	–
12	C	F		
17	C	F	*Interventionism* not covered (3)	
18	B	F		
22	-	-		Not clear (1)
27	-	-	Hard question (1) D is wordy (3)	
28	B	E		28: *la quelle → laquelle* (2)
29	-	-		29C: Delete *pour* at the end (3) Ambiguous (1)
34	-	-	*universality* not covered in class (2)	*universalité* is a hard word (2)
35	-	-		
36	C	F		
38	B	E	Hard & confusing (2)	French version is hard to understand, had to refer to the English version to answer the question correctly (1).
39	B	E	*Nostalgic* is hard (1)	B: *socialism → socialisme* (5) B: *government → gouvernement* (2)

Note. E = English, F = French. Number in brackets indicates the number of teachers who made the comment.

In regard to item clarity, four items were found confusing by the teachers. Item 1 was ambiguous for three English-speaking and two French-speaking teachers. Item 6 was confusing to two English-speaking teachers and two French-speaking teachers. One English-speaking teacher found Item 28 hard, and option D, in particular, was found wordy by three English-speaking teachers. Lastly, Item 38 was hard and confusing according to two English-speaking teachers and one French-speaking teacher. None of the above comments, however, helped explain the DIF found in Item 1, 28, or 38.

In addition to the two typographical errors discussed earlier in their contribution to DIF, two more typographical errors were identified in the French version of Items 28 and 29. It is not clear to what extent the error in Item 28 contributed to the DIF found favouring the English-speaking students. Item 29, however, did not display DIF. Besides, for Item 1, one teacher suggested to change *implantation* to *la mise en oeuvre* in French to better match *implementation* in English. The word *implantation* is often used to mean establishment and introduction.

To sum up, teacher comments identified no possible sources of DIF for mathematics, but possible explanations for four of the DIF items in social studies. Two of them appeared to be due to confusing text in French, one due to typographical errors in French, and another due to uncovered content for some English-speaking students. Although minor problems were identified in both mathematics and social studies in terms of curriculum coverage, item clarity, and linguistic correctness, with the exception of Item 28, these problems appeared to have no effect on whether an item displayed DIF or not.

Think-aloud Interviews

The purpose of the think-aloud interviews was to gain further understanding of the nature of the DIF found. An item was considered biased if it was established that the source of the extra difficulty for one group was not relevant to what the test measures (e.g., adaptation differences); otherwise the source of the DIF was considered to be undeterminable (e.g., real ability differences, curricular differences, or cultural differences) and in need of further research (Camilli & Shepard, 1994). In the case of this study, since the English-speaking students and French Immersion students attended the same schools and used the same curriculum, curricular differences were assumed to be minimal, if any. Similarly, as the English-speaking students and French Immersion students all had English as the first language and resided in English communities outside schools, the cultural differences were also expected to be negligible. Therefore, other than adaptation biases, the most likely factor contributing to DIF in this study was ability difference, or, in other words, impact.

Since monolingual English-speaking students and bilingual French Immersion students were the major intended populations for the achievement tests developed in this study, a sample of students in each of these groups were interviewed for comparison of their understanding of the test items in English and French. Francophone students were not interviewed at this time mainly for three reasons. First, Francophone students represent only 10 percent of French-speaking Grade 9 population in Alberta. Second, given the small number of Francophone students and their scattered distribution across the province, it was not possible to draw a sample of Francophone students that could be interviewed given the limited resources of the present study. Third, given the distinctiveness of the Francophone (French as the first language) population, the decision was made not to include individual Francophone students in the sample of French Immersion students (French as a second language).

Interview Sample

A sample of Grade 9 English-speaking students and a sample of Grade 9 French Immersion students were recruited from public schools in a major metropolitan area in Alberta that enrolled both English only and French Immersion students. It was made clear to the principals and teachers who assisted with the sampling that the students selected should be highly verbal in English or French and be able to think aloud while they solved the problems. Students were over-sampled to allow for non-response, denial, and absence. Altogether, 200 students (80 English and 120 French Immersion) were selected to take information letters and consent forms (see Appendix I) to their parents, who then indicated whether or not they gave consent for their children to be interviewed on an individual basis. All together 31 consent forms for English-speaking students and 44 consent forms for French Immersion students were returned. In each case, the parents gave consent for the children to be interviewed. Sixty-four students--24 English-speaking and 40 French Immersion—were interviewed, with an equal number of males and females. There were one mathematics and two social studies forms for the English-speaking students, and two mathematics forms and three social studies forms for the

French Immersion students. More French forms were needed to allow extra time for the French Immersion students to evaluate the comparability of the two language versions. For each form, eight students were interviewed.

Instruments

The instruments used in the interviews included a mathematics or social studies assessment question set, a set of structured interview guidelines (see Appendix J), and a structured observation sheet (see Appendix K).

Assessment question sets. The assessment question set for mathematics contained five items that displayed DIF and four items that did not display DIF. The corresponding numbers for the social studies question set were 17 DIF and 5 non-DIF items. All the DIF items identified by SIBTEST were included; the non-DIF items were included to mask the DIF items. The non-DIF items were selected to ensure their comparability in terms of difficulty, topic, and format to the DIF items in each question set. The number of items administered in each interview was such that each interview could be completed in one class period. To meet the requirement, the nine mathematics items were placed in one form in English and in two forms (one with five items and one with four items) in French. For social studies, the 22 items were divided into two forms (each with 11 items) in English, and three forms (two with seven items each and one with eight items) in French. The smaller number of items in the French forms allowed time for the French Immersion students to evaluate the comparability of the English version and French version of the items in the interview question set.

Interviewing guidelines. The think-aloud protocol contained a set of questions that the interviewers posed to participants upon completion of each item. Drawing on Ercikan et al. (2004), the following four questions were used with all students:

1. Are there any words that you do not know in this question?
2. In your own words, could you tell me what you believe this question is asking?
3. Did you find any parts of the question confusing? If so,
    i. What parts did you find confusing?
    ii. Why are they confusing?
4. Did you find any parts of the question helpful in solving the problem? If so,

i. What parts did you find helpful?

ii. How did they help you solve the problem?

As the usefulness of feedback from bilinguals for evaluating language equivalence has been widely recognized (Streiner & Norman, 1995), French Immersion students were also asked to compare the two language versions of each item and to look for any nonequivalence between them. After answering the four questions listed above, they were presented with the English version of the item. Upon finishing reading the English version, they were asked:

5. Do the two versions mean exactly the same thing to you? On a 3-point scale--different, similar, identical--how would you rate their comparability in meaning?

If they responded "different" or "similar", they were then asked:

6. Do you find any differences in wording? If so, where are they and how do the words differ?

Structured observation sheet. The structured observation schedule included instructions for the interviewers to record the students' gender and events that were not captured on audiotape, such as the use of gestures.

Interviewers

The researcher conducted the interviews with English-speaking students. Two bilingual interviewers were recruited to interview the French Immersion students. The first interviewer was a French native who immigrated to Canada about 10 years ago. He learned English in high school and took English courses in Canada. With a Bachelor of Education degree, he had taught French as a second language, tourism, and computer studies to junior high and high school students in Canada. The second interviewer was a Canadian Anglophone, who took a French Immersion program from Grade 7 to Grade 12. She had a Bachelor of Arts in French as well as a Bachelor of Education. She had previously taught elementary and junior high school French Immersion students and also had translation experience between English and French.

The two interviewers were asked to read and sign a confidentiality agreement (see Appendix L). The training for the interviewers took two hours, and included a review of the think-aloud procedures, structured interview guidelines, the use of the observation sheet, and the use of tape recorders and microphones.

Interview Procedure

Ericsson and Simon (1993) developed a model for verbalization processes of subjects under specific conditions so that inferences could be made about the cognitive processes that produced the verbalization. They made a distinction between two types of verbalization: concurrent and retrospective. Concurrent verbalization involves verbalizing the information one attends to while completing a task. Retrospective verbalization occurs after the task has been completed and involves recollection of one's thought processes. Retrospective reports serve to complement, elaborate, and validate the content of concurrent reports. It is recommended that, whenever appropriate, both concurrent and retrospective reports be collected (Ericsson & Simon, 1993).

In this study, both concurrent and retrospective reports were collected. Each student was trained at the beginning of his/her session with one question taken from the mathematics test if the student was to respond to mathematics items or one question taken from the social studies test if the student was to respond to social studies items. Each student was asked to talk aloud about what he/she was thinking and what information he/she was attending to while answering each question. Probes in the concurrent portion of the interview were kept to minimum. The students were only to be reminded to keep talking after 5 to 10 seconds of silence. After an answer had been selected, the students were then asked the probe questions listed above. Students who responded to the English version of the items were expected to report in English and French Immersion students who responded to the French forms were expected to report in French. However, if a French Immersions student started to think aloud in English, he/she was allowed to do so. The entire process was audio-taped.

Data Analysis

The protocols from English-speaking students were transcribed and verified by the researcher. The French Immersion students' protocols were transcribed and translated into English by the female bilingual interviewer. To check the accuracy of her translations, another bilingual person translated and transcribed a sample of the French Immersion students' responses. This person was female, born in England, educated in France, and majored in English at university. She was teaching French at the university level in Canada at the time she participated in the study. Three mathematics and seven social studies student protocols were randomly selected from the corresponding sets of French Immersion interview protocols. The number of social studies protocols selected for this part of the study was greater because social studies had three forms while mathematics had two, meaning that more students were interviewed for social studies. Further, due to the nature of the subject, social studies protocols tended to be longer and more complicated than mathematics protocols. Thus, more protocols were selected for the social studies translation verification. For each of the 10 protocols, two items were randomly selected for the second translation. The researcher compared the two versions of translation, and found a very close fit: the wording might have been different at times, but the meanings stayed the same.

Following verification of the transcripts, the protocols were interpreted and coded for each interview item. The focus of the analysis was to determine for each item: (a) how well the students understood the meaning of the question; (b) what aspects of the question, if any, hindered the students in solving the problem; (c) what aspects of the question, if any, facilitated the students in solving the problem; and (d) to what degree the two language versions were different in meaning or wording. That is, four basic categories, corresponding to each of the themes in the think-aloud protocols were created for each test question: unknown words, understanding of the question, confusing parts, and helpful parts. For the French protocols, one more category was added: comparison of the two language versions of each item.

The coding of the protocols was completed by the researcher using the coding schemes included in Appendices M and N. Students' responses to each retrospective question were examined and coded in combination with relative information from the concurrent reports. After each student's data was coded, the coding schema was revised where necessary, and the previously coded protocols were recoded using the modified classification scheme. The coding was recorded on the coding sheets (see Appendices O and P).

Reliability of the coding was then estimated by having two independent raters code a sample of the interview protocols. The two raters were experienced Grade 9 teachers from a French Immersion school. At the time of their participation in the study, one teacher was teaching mathematics and the second teacher was teaching social studies. All of the items used in the think-aloud interviews were first divided into three categories based on their difficulty level of coding: hard, medium, and easy. Then three mathematics and seven social studies items were randomly selected across the three categories. For each selected item, four student protocols were randomly selected for coding out of the eight students interviewed. That is to say, 16.7% of mathematics coded data and 15.9% of social studies coded data were reviewed by the two raters.

The training of the teachers took one and a half hours, including a review of the coding procedure and the coding schemes used in this study. One mathematics and two social studies items were selected for the purpose of training. For each item, eight transcripts (four for each language) were randomly selected. The researcher and the two raters coded the first two transcripts for each subject together. Then the two raters coded the remaining training protocols independently, followed by discussions of each transcript. Attention was paid to ensuring that the teachers had a solid understanding of the coding task they were supposed to accomplish. Following the training session, the two teachers independently coded the 12 mathematics and 28 social studies transcripts.

Inter-coder Agreement

Table 16 contains a summary of the inter-rater agreement. For the English data in both mathematics and social studies, the inter-rater agreement was above 95%. For the French data in mathematics, the agreement between the researcher (Coder 1) and the mathematics teacher (Coder 2) was 100%. The agreement between the social studies teacher (Coder 3) and the other two raters was lower (91.7%). This might be attributed to the fact that the third coder specialized in social studies, and did not fully follow students' thought process at times (based on comments by Coder 2). For the French data in social studies, the inter-rater agreements were lower (91.1% to 95.5%), but still satisfactory. This might be due to the interaction of more complex protocols and the addition of a coding theme (comparison of English and French versions). As the two teachers

Table 16

Inter-rater Agreement for the Coding of Interview Data

	Mathematics		Social Studies	
	English	French	English	French
Coder 1 vs. Coder 2	95.8%	100%	97.3%	92.9%
Coder 1 vs. Coder 3	97.9%	91.7%	95.5%	91.1%
Coder 2 vs. Coder 3	97.9%	91.7%	96.4%	95.5%
Mean	97.2%	94.5%	96.4%	93.2%

commented, coding of the mathematics data was not at all difficult, but coding of the social studies data was somewhat difficult. The mean inter-rater agreements varied from 93.2% to 97.2%. Taken together, the results revealed that the coding was reliable (Krippendorff, 1980) and valid. Of the items that did not receive 100% agreement, most of the disagreement occurred in coding students' understanding of the question. To resolve any discrepancies that occurred, discussions were held between the researcher and the two teachers until consensus was reached. Since all the discrepancies were question specific, there was no need to modify the coding and therefore no changes were made to the coding of other items that were not included in the reliability check.

Following verification of the coding, the findings from think-aloud protocols were examined and synthesized in an attempt to identify the source of DIF. For each DIF item, attention was paid to the identification of differences between students' interpretation of the two language versions of the test items, especially the differences that might account for the DIF observed. For comparison purpose, the same analysis was conducted for non-DIF items.

Interview Results for Mathematics

For the think-aloud interviews in mathematics, nine items (five DIF, four non-DIF) were administered. Eight students in each language group responded to each item. An analysis of student verbal responses suggested that no DIF could be attributed to adaptation differences.

DIF Items

For four of the five DIF items (Items 8, 9, 11, and 24), students from both language groups had little problem understanding the items, and the French Immersion students rated the two language versions the same for all the items (see Table 17). The information gleaned from the student protocols for Item 9 is presented below to illustrate this finding.

Table 17

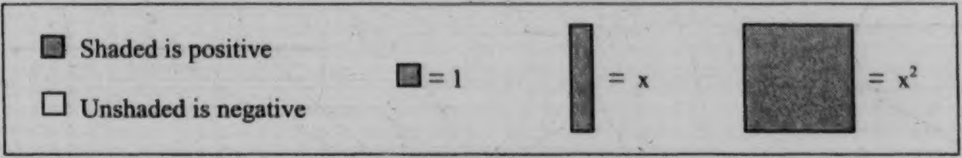
Rating of the Equivalence of Mathematics Items by French Immersion Students

Item	DIF Level	Favouring Group	Identical	Similar	Different
2	-	-	8	0	0
3	-	-	8	0	0
8	C	E	8	0	0
9	C	F	8	0	0
11	C	E	8	0	0
12	B	E	7	1	0
14	-	-	8	0	0
22	-	-	8	0	0
24	B	F	8	0	0

Note. E = English, F = French.

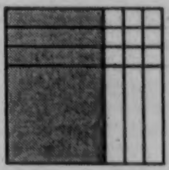
Item 9 displayed C-level DIF favouring French-speaking examinees in the field-test samples. Analysis of the student protocols suggested that the interviewed students experienced no problem understanding this item, with the exception of one French Immersion student. He thought the word *produit* meant "total answer." Besides, three French Immersion students suggested that French terminology such as *polynômes* was a bit easier to understand, which could be due to the fact that they learned mathematics in French, not in English, as one student explained. However, all of the French Immersion students indicated that the two language versions of each items were equivalent in meaning. Taken together, no language differences could be identified to have contributed to the DIF in this item. The mistake identified earlier by teachers regarding the position

Use the following algebra-tile legend to answer question 9.



▨ Shaded is positive
☐ Unshaded is negative
= 1
= x
= x^2

Use the following diagram to answer question 9.



9. The diagram above shows the product of which pair of polynomials?

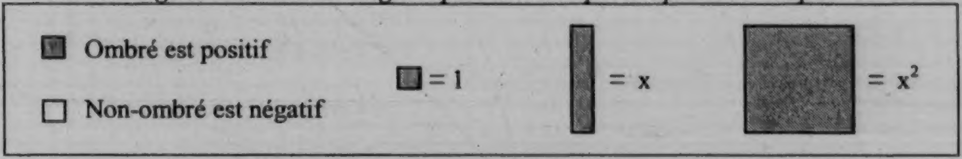
A. $(x + 3)(x + 3)$

B. $(x + 3)(x - 3)$

C. $(3x + 3)(3x - 3)$

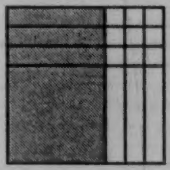
D. $(x^2 + 3x)(x^2 - 3x)$

Utilise la légende de carreaux algébriques suivante pour répondre à la question 9.



▨ Ombré est positif
☐ Non-ombré est négatif
= 1
= x
= x^2

Utilise le diagramme suivant pour répondre à la question 9.



9. Le diagramme ci-dessous montre le produit de quelle paire de polynômes?

A. $(x + 3)(x + 3)$

B. $(x + 3)(x - 3)$

C. $(3x + 3)(3x - 3)$

D. $(x^2 + 3x)(x^2 - 3x)$

of the diagram did not seem to have affected the performance of French-speaking students: the diagram was very obvious.

Item 12, however, was confusing to students from both language groups. French-Immersion students, in particular, found the English version easier to understand.

English	*French*
12. Together, three friends have $256.00. There are three times as many $5 bills as there are $20 bills and four fewer $10 bills than $5 bills. If there are three times more loonies than $5 bills, how many $20 bills are there?	**12.** Ensemble, trois amis ont 256,00 $. Il y a trois fois plus de billets de 5 $ que de billets de 20 $ et 4 billets de 10 $ de moins que de billets de 5 $. S'il y a trois fois plus de pièces de 1 $ que de billets de 5 $, combien de billets de 20 $ y a-t-il?
A. 4	A. 4
B. 8	B. 8
C. 12	C. 12
D. 16	D. 16

Item 12 displayed B-level DIF favouring English-speaking examinees. This item contained a lot of information for the students to digest. Although no students from either group identified any unknown words, four students from each language group indicated that the question was confusing because it contained too much information. These eight students either answered the question incorrectly or obtained the right answer by guessing. As one English-speaking student who answered the item incorrectly put it, "there are lots of loops and steps that you have to go through, if you want to find the answer." Of the eight students who commented about the large amount of information, two English-speaking and one French Immersion student indicated that they did not completely understand the item. They found it difficult making sense of mathematical relationships such as *three times as many... as* and *four fewer... than*.... This finding of complicated wording is similar to what teacher comments suggested earlier in this chapter.

Out of the eight French Immersion students who compared the English and French versions, seven students considered their meanings identical while one student considered them similar. Four students found the English version somewhat clearer. In particular, two of them thought *four fewer $10 bills than $5 bills* was easier and flowed better than the French counterpart *4 billets de 10 $ de moins que de billets de 5 $* [four $10 bills less than $5 bills]. One student found the English version clearer "because I read more in English and just use French at school." That is to say, although the translation of this item was correct, a high level of reading was required for both language versions. The advantage of the English group, as suggested by the exhibition of the DIF, appears to be attributed to the interaction of load of information and lack of opportunity to read everyday French on the part of French Immersion students.

Non-DIF items

To mask the DIF items, four non-DIF items (Items 2, 3, 14, and 22) were included in the question set used for think-aloud interviews. These four items all had $\hat{\beta}_{uni}$ values of close to zero. They were selected based on their correspondence to the DIF items in item format and content category. A comparison of student responses indicated no differences in the way students from the two language groups interpreted the non-DIF items. As shown in Table 17, the two language versions were perceived the same by the French Immersion students for all the four items.

Interview Results for Social Studies

For the think-aloud interviews in social studies, 22 items (17 DIF, 5 non-DIF) were administered to the students. Eight students responded to each item in either English or French with the exception of items in Form 2 (Items 3, 4, 5, 13, 25, 26, 31, and 39) of the French version. One of the eight French-speaking students interviewed using Form 2 turned out to be Francophone, so his verbal report was excluded from all subsequent analyses. Besides, for the French version of Item 39, only five verbal reports were obtained because the interviewer ran out of time.

DIF Items

An analysis of student verbal responses found no language-related evidence that helped explain the DIF in 9 of the 17 DIF items: Items 19 and 38 favoured the English group, and Items 11, 12, 13, 17, 18, 31, and 36 favoured the French group. For each of these items, all but one French Immersion students, if not all, rated the two versions identical (see Table 18). For some items, the students from the two groups found them equally easy or equally difficult to understand. For other items, some words were more familiar to one group than the other, but the differential familiarity did not affect how students answered the question. In other words, these words were either not crucial for answering the items correctly or the students were able to guess the meaning through context.

Table 18

Rating of the Equivalence of Social Studies Items by French Immersion Students

Item	DIF Level	Favouring Group	Total Number of Responses	Identical	Similar	Different
1	B	E	8	7	1	0
3	C	E	7	6	1	0
4	C	E	7	5	2	0
5	-	-	7	6	1	0
9	B	E	8	6	2	0
11	C	F	8	8	0	0
12	C	F	8	8	0	0
13	C	F	7	7	0	0
17	C	F	8	8	0	0
18	B	F	8	8	0	0
19	B	E	8	8	0	0
25	-	-	7	4	3	0
26	B	E	7	7	0	0
28	B	E	8	8	0	0
29	-	-	8	8	0	0
31	B	F	7	6	1	0
35	-	-	8	8	0	0
36	C	F	8	8	0	0
37	C	E	8	8	0	0
38	C	E	8	8	0	0
39	B	E	5	5	0	0
40	-	-	8	7	1	0

Note. E = English, F = French.

Interview data for Items 17 and 31 are presented to illustrate these findings. Item 17 represents items with vocabulary that are equally difficult for both groups, while Item 31 characterizes items with vocabulary that are more familiar to one group than the other.

Item 17 displayed C-level DIF favouring French-speaking examinees. All of the eight French Immersion students who were interviewed indicated that the two language versions meant the same, although two of them considered the English text somewhat easier to understand. Both groups shared the difficulty in understanding the correct answer interventionism/interventionnisme: four students from the English group and three from the French group did not know the word. Therefore, no evidence was found suggesting adaptation contributed to the DIF favouring French-speaking examinees on this item.

English	*French*
17. This action of the Canadian government is an example of	**17.** L'action du gouvernement canadien est un exemple
A. protectionism.	A. de protectionnisme.
B. interventionism.	B. d'interventionnisme.
C. partnership with the private sector.	C. de partenariat avec le secteur privé.
D. cooperation with the private sector.	D. de coopération avec le secteur privé.

Unlike Item 17, differential familiarity with vocabulary was found between the two language groups in Item 31, which displayed B-level DIF favouring French-speaking examinees. Six of the seven French Immersion students interviewed indicated that the two versions were identical in meaning, while one student considered them similar. Two students considered the English version somewhat easier because of the word *supporter/ partisan.* No English-speaking students reported difficulty in understanding *supporter*. In contrast, six French Immersion students did not understand *partisan.* This difference, however, did not seem to have hindered the French Immersion students from answering the question correctly. As two students explained, they simply interpreted *partisan* as person, a person living in a mixed economic system. Therefore, no language differences were found explaining the DIF favouring French-speaking examinees in this item.

English	*French*
31. A supporter of a mixed economic system would say that the **most** important indicator of quality of life is	**31.** Un partisan de l'économie mixte dirait que l'indicateur **les plus** important de la qualité de vie est
A. guarantee of a job for life.	A. la garantie d'un emploi pour la vie.
B. access to social services.	B. l'accès aux services sociaux.
C. protection of the environment.	C. la protection de l'environnement.
D. choice of consumer goods.	D. le choix de biens de consommation.

Some evidence was found explaining the DIF in 8 of the 17 DIF items: Items 1, 3, 4, 9, 26, 28, 37, and 39. Interestingly, all these items favoured the English group. In terms of possible causes for DIF, these items can be classified into two groups: differential familiarity with key words/concepts (Items 1, 9, 26, and 37), and differential difficulty of stimulus texts (3, 4, 28, and 39).

Differential familiarity with key words/concepts. Item 1 displayed B-level DIF favouring English-speaking examinees.

English	*French*
1. What was the impact of the implementation of mass production on the workers?	**1.** Quel effet a eu l'implantation de la production en série sur les travailleurs?
The specialization of work and	La spécialisation du travail et
A. a reduction in working hours.	A. la réduction des heures de travail.
B. improved working conditions.	B. l'amélioration des conditions de travail.
C. more control over production.	C. plus de contrôle sur la production.
D. loss of control over the end product.	D. la perte de contrôle sur le produit final.

Seven of the eight French Immersion students found the two language versions the same in meaning, while one student found them similar in meaning. Four English-speaking students did not understand the word *implementation*, and five French Immersion students had trouble interpreting the French equivalent, *l'implantation*. While the words *implementation/ l'implantation* caused trouble for about the same number of students from both groups, another key concept, *mass production,* was found difficult only by the French Immersion students. Three of the eight students did not know the French equivalent, *production en série,* and these three students answered this item incorrectly. Two English-speaking students acknowledged the phrase *mass production* as helpful in determining the right answer. Therefore, although the two phrases mean the same, *mass production* seems to be easier for the English group than *production en série* for the French group. In other words, their differential familiarity with this key concept might have led to B-level DIF favouring the English group.

Similarly, Item 9 displayed B-level DIF favouring English-speaking examinees.

English	*French*
9. Which of the following revolutionary practices allowed Henry Ford to produce a good quality car at an affordable price?	**9.** Laquelle de ces pratiques révolutionnaires a permis à Henry Ford de créer une voiture de bonne qualité à un prix modéré?
A. Formation of monopolies	A. La formation de monopoles
B. Introduction of closed shops	B. L'introduction d'ateliers fermés
C. Invention of the assembly line	C. L'invention de la chaîne de montage
D. Cooperation of the trade unions	D. La coopération des syndicats

Six of the eight French Immersion students found the two versions the same in meaning, while the other two found them similar in meaning. Three French Immersion students had trouble making sense of *la chaîne de montage*, while no English-speaking students had difficulty understanding the English counterpart, *assembly line*. The phrase *ateliers fermés* was found confusing by four French Immersion students, while no English-speaking students had difficulty with the English counterpart *closed shops*. Two students from each group had trouble with the word *monopolies/monopoles*. Taken together, students' differential familiarity with the key concept *assembly line* as well as the phrase *closed shop* might have led to the exhibition of B-level DIF favouring the English group. Although the two versions mean the same, the French version seems to be more difficult for the French Immersion students to understand due to their lack of knowledge of key words.

Item 26 was one of the three items that were referenced to the same text (see Appendices G and H). While the other two items (Items 24 and 25) referenced to the text did not exhibit DIF, Item 26 displayed B-level DIF favouring English-speaking examinees. For comparison purpose, Item 25 was also administered in the think-aloud interviews.

English	*French*
25. Who demonstrates the **greatest** support for a free market economy? A. Jean B. Michelle C. Paul D. Alexandra	**25**. Qui manifeste **le plus** d'appui pour une économie de marché libre? A. Jean B. Michelle C. Paul D. Alexandra
26. Who expresses the **most** concern about the issue of unemployment? A. Jean B. Michelle C. Paul D. Alexandra	**26**. Qui manifeste **le plus** d'inquiétude pour le problème du chômage? A. Jean B. Michelle C. Paul D. Alexandra

The seven French Immersion students interviewed found the two versions of the text identical in meaning. For item 25, one French Immersion student reported she did not understand the word *manifeste* [demonstrate], but this lack of understanding did not seem to prevent her from interpreting what the question is asking. As she explained, "I am not sure what the word *manifeste* means, but it [the item] wants to know who was supportive of a market economy."

For Item 26, all the seven French Immersion students who were interviewed considered the two language versions the same in meaning. However, one student did not know the key word *chômage* [unemployment], and two others expressed the concern that some of their classmates might not be familiar with this word. Although the translation was correct for this item, the students' differential familiarity with the key word might to a certain degree contributed to the B-level DIF favouring the English group.

Lastly, Item 37 displayed C-level DIF favouring English-speaking examinees.

English	*French*
It is not what strangers think of the Perestroïka that is important, say the Soviets, what matters is what happens here. Within the last 5 years, instead of improving, the situation has become worse. The grocery stores offer fewer products and the stores, with their poor quality and their old-fashioned clothes, look more and more like the Salvation Army store.	Ce n'est pas ce que les étrangers pensent de la Perestroïka qui est important, disent les Soviétiques, c'est ce qui se passe ici. Depuis 5 ans, loin de s'améliorer, la situation économique du pays s'est aggravée. Les comptoirs d'alimentation offrent de moins en moins de produits et les magasins, avec leur marchandise de mauvaise qualité et leurs vêtements démodés, ressemblent de plus en plus à des comptoirs de l'Armée du Salut.
37. What is the problem identified in this paragraph?	**37.** Quel est le problème évoqué dans ce paragraphe?
A. Scarcity	A. La pénurie
B. Repression	B. La répression
C. Corruption	C. La corruption
D. The black market	D. Le marché noir

All of the eight French Immersion students who were interviewed reported that the two language versions meant the same, although seven of them found the English text clearer. No English-speaking students reported problems understanding the key answer *scarcity*, while five French Immersion students had trouble understanding the French counterpart, *la pénurie.* These five French Immersion students all considered the English version somewhat clearer. Two mentioned specifically that they recognized *scarcity* in English but not *la pénurie* in French. In addition, two students switched their answer to A, the correct answer, after reading the English version. That is to say, although the two words mean the same, *scarcity* seems to be an easier word for the English group than *la* *pénurie* for the French group. In other words, their differential familiarity with the key answer might have led to the exhibition of C-level DIF favouring the English group.

To sum up, the four items presented above shared one thing in common: the French Immersion students were not as familiar with some key words or concepts as their English counterparts. These key words and concepts were crucial, however, in determining the best answer. Therefore, this differential knowledge might have contributed to the occurrence of DIF in these items favouring the English group.

Differential difficulty of stimulus texts. Items 3 and 4 are two of the three items that were referenced to the same stimulus text (see Appendices G and H). They both displayed C-level DIF favouring English-speaking examinees. The third item (Item 5) did not exhibit DIF. For comparison purpose, Item 5 was also administered in the think-aloud interviews. With the Francophone student excluded, seven French protocols were used in the following analysis.

The text used for Items 3, 4, and 5 is one of the longer texts included in the social studies test. The English version is 153 words long and the French version is 189 words long. In terms of the comparability of the two versions of the text, six out of the seven French Immersion students said that the English version was clearer/easier than the French version. The reasons underlying their judgments included: "the terms used in French were quite difficult to understand"; "I recognize lots of the terms in English"; and "The words [in English] were more familiar to me and I really don't like reading in French". While only one of the eight English-speaking students found the text confusing, four of the seven French Immersion students found it hard to understand. When it comes to unknown words, two English-speaking students mentioned the word *municipal*, which they did not think affected how they answered the questions. In contrast, two French Immersion students did not understand *égout* [sewer], which happened to be an important word for understanding the messages conveyed through this text.

For Item 3, six of the seven French Immersion students considered the two versions identical in meaning, and one considered them similar. While no English-speaking students had difficulty understanding this item, two French Immersion students reported problems with the key word *appuie* [support]. Therefore, students' differential

English	*French*
3. According to this text, we can say that CUPE Local 287	3. Selon ce texte, on peut affirmer que la section locale 287 du SCFP
A. supports the North Battleford city council.	A. appuie le conseil de ville de North Battleford.
B. supports the privatization of the sewer treatment plant.	B. appuie la privatisation de la station d'épuration des eaux d'égouts.
C. opposes the privatization of the sewer treatment plant.	C. s'oppose à la privatisation de la station d'épuration des eaux d'égouts.
D. supports a partnership with public and private sectors in the issue of sewer treatment.	D. appuie un partenariat public-privé dans le dossier de d'épuration des eaux d'égouts.
4. According to this text, what is the **most** important issue raised by CUPE Local 287?	4. Selon ce texte, quelle est la préoccupation **principale** de la section locale 287 du SCFP?
U.S. Filter Canada	U.S. Filter Canada
A. will not do a good job.	A. ne fera pas un bon travail.
B. is a property of a French multinational.	B. est la propriété d'une multinationale française.
C. is more interested in profit than water quality.	C. est plus intéressée par les profits que par la qualité de l'eau.
D. threatens the job security of the municipal workers.	D. est une menace pour les emplois des employés municipaux.
5. According to this text, CUPE Local 287 is a	5. Selon ce texte, la section locale 287 du SCFP est
A. group of concerned citizens from North Battleford.	A. un groupe de citoyens inquiets de North Battleford.
B. business specialized in sewage treatment.	B. une entreprise d'épuration des eaux d'égout.
C. non-profit organization.	C. une association à but non-lucratif.
D. workers' union.	D. un syndicat de travailleurs.

understanding of this key word and the text in general might have caused the differential performance favouring the English group.

For Item 4, five French Immersion students considered the English and French versions identical, while two considered them similar because the most important issue and la préoccupation principale had different emphases. To them, la préoccupation principale meant main concern rather than the most important issue. It is not clear whether this difference affected how students from the two groups selected their answer to this question. However, one thing in common for Items 3 and 4 is that they both required adequate understanding of the whole text. For Item 3, eight English-speaking students and three French Immersion students found Paragraph 3, especially the last part, helpful in answering the question. One French Immersion student acknowledged Paragraph 1 helpful. Similarly, for Item 4, seven English-speaking students and four French Immersion students found Paragraph 3 helpful. Therefore, the DIF for Items 3 and 4 might have been caused by the differential difficulty of the text, because answering these items involved sufficient understanding of the whole text.

In contrast to Items 3 and 4, Item 5 did not display DIF. Six of the seven French Immersion students considered the two versions the same in meaning, while one considered them similar. Analysis of student verbal reports indicated that the word *non-lucratif* [non-profit] in the text was a new word for all the seven students in the French group, while students from the English group had no problem at all understanding this item. In spite of this difference, Item 5 did not produce DIF. For one thing, *non-lucratif* was not in the right option. For another, answering Item 5 required access to local information rather than global information, as compared to Items 3 and 4. All the eight English-speaking students and three of the French Immersion students found one sentence particularly helpful in answering this question: "CUPE Local 287, which represents 123 municipal workers including sewer and water plant operators, outlined their concerns …." In other words, this definition of CUPE Local 287 provided key information for answering Item 5. Therefore, student performance on Item 5 might not have been as much affected by the differential difficulty of the text as their performance on Items 3 and 4.

Similar to Items 3, 4, and 5, Items 27, 28 and 29 were reference to the same text (see Appendices G and H), but only Item 28 displayed B-level DIF favouring the English group.

English	*French*
28. Marlene Sandberg's business was successful for which one of the following reasons?	**28.** L'entreprise de Marlene Sandberg a eu du succès pour la quelle des raisons suivantes?
A. She competed with large companies.	A. Elle faisait concurrence à de grandes compagnies.
B. She marketed her product very well.	B. Elle a bien commercialisé son produit.
C. She solved the issue of pollution in Sweden.	C. Elle a réglé le problème de la pollution en Suède.
D. She created a product more efficient than the original disposable diapers.	D. Elle a créé un produit plus efficace que les premières couches jetables.
29. According to the law of supply and demand, what factor is the **supply** in Marlene Sandberg's story?	**29.** Selon la loi de l'offre et de la demande, quel facteur représente **l'offre** dans l'histoire de Marlene Sandberg?
A. Disposable diapers	A. Les couches jetables
B. Biodegradable diapers	B. Les couches biodégradables
C. Children's need for diapers	C. Le besoin des enfants pour des couches pour
D. People's concern about the environment	D. La préoccupation du public pour l'environnement

The stimulus text was found easier in English by six of the eight French Immersion students. In particular, four students pointed out that they did not know the word *couches* [diapers], which was a key word in understanding the whole passage. Two of them also did not understand another key word, *jetables* [disposable]. In contrast, no English-speaking students reported any problem understanding the stimulus text.

For Item 28, all eight French Immersion students deemed the two language versions identical in meaning, but two students did not know that the word *concurrence* in option A meant *competition* in English. This difference in combination with the differential difficulty of the text might have contributed to the DIF in Item 28 favouring the English group. For comparison purpose, Item 29, which did not show DIF, was included in the interview. All eight French Immersion students considered the two versions identical in meaning, although one of them found the English version easier because he did not know *préoccupation* [concern] in option D. It is not clear, however, why this item and Item 27 which were based on the same text, were DIF free.

Lastly, Item 39, which was referenced to a picture with an accompanying text, displayed B-level DIF favouring English-speaking examinees. The French version text contained 67 words, while the length of the English version text was 49 words. Eight English-speaking students completed the think-aloud reports, but only five French Immersion students provided their verbal reports because the interviewer was out of time. Four of the five French Immersion students found the English version of the text clearer, while the other student found the French version easier (e.g., words such as *manifestants* [protesters] and *humiliation* [humiliation]). Three of the four students switched their answer to B, the correct answer, after reading the English version of the text and question. As one student explained, "I understood the French version, but the English text was easier because it's my first language. Also the second answer [B] is a lot more straight-forward in English." Another student echoed, "I guess they [the two language versions] shared the same message, but I found the English one shorter and less complicated. In French they used some vocabulary that you really have to think about." Therefore, the DIF for Item 39 might have been caused by the differential difficulty of the stimulus text, because answering this item involved adequate understanding of the whole text.

To sum up, for the six items discussed above, the stimulus text was more difficult for the French Immersion students than for the English-speaking students. As a result, this differential difficulty might have contributed to the occurrence of DIF favouring the English group on four of the six items. However, at the same time, other items related to the same text did not display DIF. For one item, this might be due to the nature of the questions that required adequate understanding of specific parts of the text rather than the whole text. It was not clear why the other items that were referenced to the same text as a DIF item did not display DIF.

English	*French*
Moscow Communists are Nostalgic	Les Communistes se sentent nostalgiques à Moscou
Pro-communist protesters shout anti-government slogans as they rally holding the Soviet hammer and sickle flag to mark the 82nd anniversary of the Bolshevik Revolution in downtown Moscow on Sunday. The Russian Communist Party on Saturday said that the government's rejection of socialism had resulted in "an unprecedented national humiliation."	Des manifestants en faveur du communisme crient des slogans contre le gouvernement alors qu'ils manifestent en tenant le drapeau soviétique du marteau et de la faucille pour marquer le 82e anniversaire de le révolution bolchévique dans aucentre ville de Moscou dimanche. Le parti communiste russe a déclaré samedi que le rejet du socialisme par le gouvernement avait eu comme résultat « une humiliation nationale sans précédent ».
39. The **best** conclusion that can be drawn from the newspaper article is that	**39.** La **meilleure** conclusion que l'on peut tirer de l'article de journal est que
A. the public will never support the Russian government as long as it rejects communism.	A. le gouvernement russe n'aura pas l'appui du public tant qu'il rejettera le communisme.
B. the government's rejection of socialism created uncertainty for some Russian citizens.	B. le rejet du socialism par le governement représente une incertitude pour des citoyens russes.
C. the Russian government made a mistake when it adopted capitalism.	C. le gouvernement russe a fait une erreur en adoptant le capitalisme.
D. no political system can ever satisfy the public.	D. aucun système politique ne peut satisfaire le public.

Non-DIF items

To mask the DIF items, five non-DIF items were included in the question set used for think-aloud interviews. These five items all had $\hat{\beta}_{uni}$ values of close to zero, and were selected based on their correspondence in topic and format to the DIF items. Four of these non-DIF items (Items 5, 25, 29, and 35) shared the same stimulus text with at least one DIF item, and one (Item 40) was similar in topic and format to a DIF item (Item 39). Items 5, 25, and 29 were examined earlier together with the DIF items (Items 3, 4, 26, and 28). Items 35 and 40 are discussed next.

Item 35 shared a stimulus material with Item 36 that displayed C-level DIF. The stimulus texts in the two languages were found to mean the same by seven of the eight French Immersion students. The other student found them similar, and the English version easier to understand. For both Items 35 and 36, all eight French Immersion students perceived the two versions the same in meaning. Four students from each group did not know the key word *universality/universalité* in Item 35, which made it not surprising that this item functioned equally for the two groups. Item 36, however, displayed C-level DIF favouring the French group although no students from either group had difficulty understanding the question. No supporting evidence in regard to language differences was found for the DIF on Item 36.

Item 40 was similar to Item 39 in topic and format. They both contained cartoons in the stimulus material, although Item 39 also included a text. As discussed earlier, the DIF of Item 39 might have been caused by the differential difficulty of the text for the two groups. Item 40, however, was found to mean the same in both versions by all but one French Immersion students, who considered the two versions similar and the options in English somewhat clearer. The word *pénurie* [shortage/scarcity] was found difficult to understand by two French Immersion students, which validated to a certain degree the finding in Item 37 that five out of the eight French Immersion students did not understand the correct option *pénurie*. For Item 40, however, this difference did not occur in the correct option, which could be why Item 40 was DIF free.

CHAPTER SEVEN

DISCUSSIONS AND CONCLUSIONS

This chapter is organized in four sections. In the first section, the purpose of the study and the research questions are outlined. The second section offers a summary of the methods and a discussion of the key findings. The efficacy of the simultaneous test development approach is addressed in the third section. The final section contains a discussion of the limitations and implications for future research.

Summary of Research Questions

This study was part of a large-scale research project designed to investigate the validity and utility of the simultaneous approach to the development of equivalent achievement tests in French and English. The major objectives of the large-scale project were to a) develop Grade 9 mathematics and social studies tests in French and English employing the simultaneous approach, b) validate the tests produced, and c) evaluate the utility of the simultaneous approach in terms of its efficiency and effectiveness. The first stage of the research project, which involved test development, item revision, and pilot testing, was reported in Rogers et al. (2003) and was reviewed in Chapter 3. The pilot test results revealed that the French-speaking examinees outperformed the English-speaking examinees in both mathematics and social studies. What was not clear is why this difference occurred. Possible reasons include non-equivalence of the tests constructed in the two languages, the presence of socio-economic differences between the two language groups (i.e., real differences in ability), or a combination of both (Rogers et al., 2003).

Consequently, the purpose of this study, which comprised the second, third, and fourth stages of the research project, was to disentangle these two issues to obtain a clearer view of the efficacy of the simultaneous approach in reducing construct bias and enhancing linguistic and cultural decentering. In particular, the following research questions were addressed:

1. How comparable are the English and French versions of the Grade 9 achievement tests in mathematics and social studies constructed using the simultaneous test development approach?

2. Is there evidence of differential item performance for English- and French-speaking examinees on the above tests?
3. If so, to what degree is the source of differential item performance related to adaptation differences?

Question 1 was addressed employing evidence from judgmental review, item analysis and factor analysis. Question 2 was addressed by analysis of differential item functioning (DIF). Teacher comments and students' interview data were used to address Question 3. Unlike previous applications of the simultaneous test development (e.g., Erkut et al., 1999; Solano-Flores et al., 2002), this study utilized comprehensive data from both empirical and substantive sources to evaluate the equivalence of the bilingual tests developed.

Discussion

Question 1: How comparable are the English and French versions of the Grade 9 achievement tests in mathematics and social studies constructed using the simultaneous test development approach?

Question 1 was addressed employing judgmental review, item analysis, and evaluation of factor structures. After the item writers completed the revision of the 58 mathematics items and 51 social studies items retained from Stage 1, certified translators reviewed these items for comparability in meaning and wording. Then three members of the research team completed the revision based on the item review results. Next, one test of 28 mathematics items and one test of 40 social studies items were constructed in both English and French. Item analysis was then conducted on the field test data of these items using the LERTAP item analysis program (Nelson, 2000), and the structural equivalence of the two language versions of the tests were evaluated using non-linear factor analysis and multi-group confirmatory factor analysis.

Mathematics. Based on the review of five translators on mathematics items, 91.4% of the items were identical in meaning in English and French, 1.7% of the items had similar meaning, while 6.9% of the items were considered different in meaning. Among the four items considered different, three were due to typographical errors. In terms of their comparability in wording, four translators marked 5.8% of the possible cells (58 items × 4 wording categories × 4 raters). Therefore, language differences between the English and French versions of the mathematics test were marginal. Accordingly, the follow-up revisions were minor.

Based on the field test data, the French-speaking examinees performed significantly better than the English-speaking examinees (with mean scores of 15.04 vs. 13.14). This finding is consistent with what was found in Alberta Grade 9 provincial achievement tests (Gierl & Khaliq, 2001). The standard deviation, kurtosis, and skewness were similar between the two language groups. The internal consistencies and average item discriminations were almost identical for the two groups. The average item difficulty for the French group was higher than that for the English group, corresponding to the fact that the French group outperformed the English group. Taken together, the psychometric characteristics of the English and French versions of the mathematics test were comparable, with the one exception being that the French group outperformed the English group.

The structural equivalence of the two language versions of the mathematics test was then evaluated using NOHARM (Fraser, 1988) for non-linear factor analysis and LISREL 8.14 (Jõreskog & Sõrbom, 1996) for multi-group confirmatory factor analysis. Item-level non-linear factor analysis suggested that one-factor model fitted the data well for both English and French groups. The parcel-level multi-group confirmatory factor analysis confirmed that English and French data were unidimensional with equal factor loadings and error variances.

Based on the judgmental review, item analysis, and factor analysis, the English and French versions of the mathematics test were comparable in terms of meaning, psychometric characteristics, and factor structures.

Social studies. According to the review of five translators on the social studies items, 64.7% of the items contained identical meaning in English and French, 29.4% of the items had similar meaning, while 5.9% of the items were considered different in meaning. One of the three items considered different was due to typographical errors. In terms of their comparability in wording, four translators marked 16.8% of the possible cells (51 items × 4 wording categories × 4 raters). Considering the content nature of social studies, language differences between the English and French versions of the social studies test were not major. Accordingly, the follow-up revisions were not substantial.

Based on the field test data, the French-speaking examinees performed significantly better than the English-speaking examinees (with mean scores of 27.75 vs. 25.74). This finding is consistent with what was found in Alberta Grade 9 provincial social studies achievement tests (Gierl & Khaliq, 2001; Alberta Learning, 2002a). The standard deviation, skewness, kurtosis, and internal consistencies were comparable between the two language groups. The average item difficulty for the French group was higher than that for the English group, corresponding to the fact that the French group outperformed the English group. Consistent with what Gierl and Khaliq (2001) found, the mean discrimination index was somewhat lower on the test written by the French-speaking examinees than the test written by the English-speaking examinees, but was still comparable across language groups. Taken together, the psychometric characteristics of the English and French versions of the social studies test were comparable, with the one exception being that the French group outperformed the English group.

The structural equivalence of the two language versions of the social studies test was then evaluated using NOHARM (Fraser, 1988) and LISREL 8.14 (Jöreskog & Sörbom, 1996). Item-level non-linear factor analysis suggested that one-factor model fitted the data well for both English and French groups. The parcel-level multi-group confirmatory factor analysis confirmed that English and French data were unidimensional with equal factor loadings and error variances.

To summarize, based on judgmental review, item analysis, and factor analysis, the English and French versions of the social studies test are comparable in terms of meaning, psychometric statistics, and factor structures. The only exception is that the French-speaking students outperformed the English-speaking students in both mathematics and social studies, which has also been typically observed in Alberta provincial achievement tests (Gierl & Khaliq, 2001; Alberta Learning, 2002a). The superior performance of the French-speaking students, 90% of whom were French-Immersion students, could be due to many factors. Other than the factor of inadequate adaptation under study in the current research, three more factors might have contributed to this finding. First, French Immersion students in Alberta tend to be from families of higher socio-economic status, and according to the PISA English reading study (Allen, 2004), differences in family socio-economic background contributed significantly to the high reading achievement of students in French immersion programs. Allen (2004) has suggested, however, that gender, socio-economic background, and parents' education did not explain all the differences between the differential performance of the English and French groups. One other factor that might have contributed to the difference is self-selection. There may be a tendency for less-skilled students not to enter French-immersion programs, or to transfer out of immersion programs if there is a concern about their ability to learn in the second language (Allen, 2004). Besides, the higher academic success of the French Immersion students might be attributed to an enriched learning environment offered in the French Immersion programs or simply the effect of bilingualism (Allen, 2004). It is also quite possible that all the factors discussed above were at play.

Question 2: Is there evidence of differential item performance for English- and French-speaking examinees on the tests?

To answer this question, Differential Item Functioning (DIF) analysis was conducted using SIBTEST to identify items that functioned differentially for the English- and French-speaking examinees. The guidelines proposed by Roussos and Stout (1996) were used to classify DIF items. Items with moderate or high level ratings, $\left|\hat{\beta}_{uni}\right| \geq 0.059$ ($p < 0.05$), were considered DIF items, while those with negligible ratings, $\left|\hat{\beta}_{uni}\right| < 0.059$ ($p < 0.05$), were not considered DIF items.

Mathematics. Five of the 26 (19.2%) mathematics items displayed DIF, with three favouring the English group and two favouring the French group. That is, the DIF items were approximately evenly distributed between the language groups. Compared with prior DIF studies of adapted tests for Canadian English- and French-speaking students, the proportion of DIF items identified in this study (19.2%) was comparable. For example, using SIBTEST, Gierl, Jodoin, and Ackerman (1999) identified DIF on 14.0% of the items on a Grade 6 mathematics provincial achievement test, and 18.4% of the items on a Grade 9 mathematics achievement test. Similarly, Ercikan (2002) found 14.1% of DIF items on TIMSS mathematics tests using an IRT-based L-H method. In another study by Ercikan et al. (2004), DIF was found on 38% and 33% of the items on a School Achievement Indicators Program (SAIP) mathematics test for 13-year-olds and 15-year-old respectively. This study also used the IRT-based L-H method to detect DIF.

Social studies. Of the 40 social studies items, 17 (42.5%) were identified with DIF, with 10 favouring the English group and seven favouring the French group. That is, the number of DIF items favouring the English group was somewhat larger than the number favouring the French group. Compared with prior DIF studies of adapted tests in English and French, the proportion of DIF items identified in this study (42.5%) was somewhere between the high and low. Gierl and Khaliq (2001), for example, identified DIF on 58.0% of the items on a Grade 6 social studies provincial achievement test, and 30.9% of the items on a Grade 9 social studies achievement test. Their study used the same DIF detection procedure, SIBTEST, as the present study.

To sum up, a larger percentage of DIF was found in social studies than in mathematics. Compared with previous research (Gierl and Khaliq, 2001; Ercikan, 2002; Ercikan et al., 2004), the proportions of DIF found in both mathematics and social studies were in the middle range.

Question 3: To what degree is the source of differential item performance related to adaptation differences?

Teacher comments and students' interview protocols were used to address Question 3. The teachers whose classes completed the field tests were asked to comment on the test items, in particular, to identify problems with the items and topics they had not yet covered. Their comments were then synthesized to identify possible sources of DIF. For the same purpose, concurrent and retrospective verbal reports (Ericsson & Simon, 1993) were collected from a sample of 24 English-speaking and 39 French Immersion students. Responses of examinees from the think-aloud interviews were used to determine whether the examinees were helped or hindered by certain aspects of the items. Similar to what Ercikan et al. (2004) established, the think-aloud protocol approach was found to be useful in identifying sources of DIF as a complementary tool to approaches like judgmental reviews and statistical methods.

Mathematics. Fifteen English-speaking teachers and 10 French-speaking teachers provided comments on the English and French versions of the mathematics test, respectively. The comments, however, did not provide obvious clues for the five DIF items. The DIF item that received the most comments was Item 12, the wording of which was considered complicated by 10 English-speaking teachers and one French-speaking teachers. Nevertheless, this item favoured English-speaking students. Teacher comments did not help explain the DIF in this item.

No "student" support for adaptation as a source of DIF was identified in any of the five DIF items in mathematics. For four (Items 8, 9, 11, and 24) of the five DIF items, no students from either language group had difficulty understanding the items, and the French Immersion students indicated that the two versions were identical in meaning. For Item 12, however, four students from each language group found the question confusing because it contained what they believed to be too much information. Four French Immersion students, in particular, found the English version easier to understand. The DIF in this item, therefore, appears to be attributed to the interaction of a heavy load of information and inadequate proficiency in French for French Immersion students.

Social studies. Eighteen English-speaking teachers and nine French-speaking teachers provided comments on the English and French versions of the social studies test, respectively. The comments helped to explain the DIF in four of the 17 DIF items. For Items 3 and 4, while the English-speaking teachers had no comments about the texts or questions, two French-speaking teachers found the third paragraph of the stimulus text confusing. This provided some clues as to why the items favoured the English group. For Item 17, three English-speaking teachers indicated that they had not covered the concept of interventionism, which was the right answer. No French-speaking teachers made similar comments. The difference in the two sets of comments may help to explain why this item favoured the French group. For Item 39, the two typographical errors identified in the correct option might have contributed to the DIF.

Protocol analysis of the student interview data revealed no evidence to explain the DIF in nine items: Items 19 and 38 which favoured the English group, and Items 11, 12, 13, 17, 18, 31, and 36 favoured the French group. For all these items except Items 13 and 31, all the French Immersion students rated the two versions identical in meaning (see Table 18). Items 13 and 31 were rated identical by all but one French Immersion student. For some items, the students from the two groups found them equally easy or equally difficult to understand. For other items, selected words were more familiar to one group than the other, but the differential familiarity did not affect how students answered the question. In other words, these words were either not crucial for answering the items correctly or the students were able to determine the meaning through context. Therefore, the DIF in these items could be attributed to ability differences or other unknown factors.

For the remaining eight DIF items, the DIF could be attributed to either differential familiarity with key words/concepts or differential difficulty of stimulus texts. As discussed earlier, some key words/concepts in the social studies items were found to be more difficult for French Immersion students. These words/concepts included mass production (Item 1), assembly line (Item 9), unemployment (Item 26), and scarcity (Item 37). Based on the judgmental review and further consultation with the bilingual team member of this project, the translations of these key words/concepts were correct. An examination of Alberta Grade 9 social studies provincial tests (1998-2001) revealed that these key words/concepts were part of these tests as well. The corresponding translations of these key words were the same as those used in this study.

In addition, some stimulus texts were also found to be more difficult for the French Immersion students. The text for Items 3, 4 and 5 served as a good example. On one hand, due to the nature of the languages, the French text was longer than the English version. On the other hand, a great majority of the French Immersion students (six out of seven) found the English version clearer and easier. As one student summarized, "the English one [version] was shorter and less complicated. In French they used some vocabulary that you really have to think about." Therefore, the differential difficulty of the stimulus texts might have contributed to the DIF found in two of the three items referenced to the same text. Similarly, the texts for Items 28 and 39 were found easier to understand by all but one French Immersion students, which could have caused the DIF in these two items.

Adaptation Bias or not? Before addressing the issue of adaptation bias, it is crucial to understand the cause underlying the two groups' differential difficulty with the key words or texts. A review of literature on French Immersion education suggested that the differential difficulty of these words/texts could be attributed to French Immersion students' lack of exposure in French outside the classroom. Day and Shapson (1996) provided a lucid example:

> A grade 4 item asking students what they should do first if a piece of bread were caught in a toaster illustrates problems in using translated items for different groups of students. The English version of this item contains words that are familiar to English-speaking children (e.g., toaster, plug, poke); however, the corresponding words in French may not necessarily be known by immersion students because their experiences in French tend to be limited to the classroom. Even though the translation is correct, the items are not equally difficult in the two languages because of the different linguistic experiences of the two groups (p. 16).

Similarly, Romney, Romney, and Braun (1989) found that French Immersion students' knowledge of words related to out-of-school activities was limited and this impeded their reading considerably. In their study of Grade 5 French Immersion students in Alberta, Romney et al. (1995) indicated that the main difficulty for French Immersion students to read in French was vocabulary. Further, Romney et al. also found that more than two-thirds of the French Immersion students never read at all in French for pleasure outside school: the average amount of time French Immersion students devoted to outside reading in French was 25 minutes a week compared to 183 minutes a week in reading in English. Likewise, they watched considerably less television in French (8 minutes per week) than in English (478 minutes per week).

In addition, French Immersions students' disadvantage in writing French tests as compared to English tests in language intensive areas has been demonstrated in earlier research. As illustrated in the study by Morrison and Pawley (1983), Grade 10 French Immersion students performed less well in history when they were tested in French than when they were tested in English. Similarly, Samuel (1990) documented that French Immersion students were not as able to demonstrate their knowledge and skills in social studies when tested in French as when they were tested in English. When French and English forms of social studies achievement tests were randomly assigned to Grade 6 French Immersion students in Alberta, those who wrote in French achieved significantly lower scores than those who took the English test. In particular, the effect sizes on topic specific text-based questions were all larger than the effect sizes on the same topic single discrete items. In another three-year study conducted in Alberta, Grades 3 and 6 French Immersion students were found to perform at a lower level when they wrote science, mathematics, and social studies tests in French than when they wrote them in English. The same pattern was also found in Grade 9 social studies tests (Alberta Education, 1990, 1991, & 1992). Taken together, these studies indicated that testing in French may underestimate their subject matter knowledge, especially in language-intensive subjects like history and social studies. These studies provided supporting evidence for the contribution of French Immersion students' inadequate proficiency in French to the exhibition of DIF favouring English-speaking students found in the present study.

An item is considered biased if it is established that the source of the extra difficulty for one group is not relevant to what the test measures (Camilli and Shepard, 1994). Adaptation bias could be introduced when the vocabulary, sentences, or texts do not have equivalent meaning or they have equivalent meaning, but are more difficult for one language group than another. Nevertheless, the extra difficulty for French Immersion students in this study appears to be caused by their inadequate proficiency in French rather than inadequate adaptation. First, the test items were written by French Immersion program teachers who were teaching the subject for which they developed items. They were familiar with the curriculum contents as well as the language levels of their students, so that the language they used should approximate the language level of their students. Second, the vocabulary and texts used in the test items were judged to be generally comparable by the certified translators. Third, most of the French Immersion students (more than half in most cases) had no problems understanding the items in French, which means that their French proficiency had reached a level they were expected to achieve.

Implications for Test Administration to French Immersion Students

For the key words/concepts found to be differentially difficult in this study, not much could be done to correct the problem, especially for concepts like mass production and assembly line. For the students who had not attained the level they were expected to achieve, attention should instead be paid to increasing their exposure in French both in and out of schools. Emphasis should also be placed on making sure they master the key concepts or words (e.g., assembly line) in French, as one of the coders in this study suggested.

Alternatively, accommodations can be made for the French Immersions students when they take tests in French. First, French Immersion students should be allowed to use a dictionary when writing tests in French. Second, as Day and Shapson (1996) indicated, visual cues and glossaries can be provided for obscure or technical vocabulary found to be difficult in the field tests. Third, from the perspective of test administration, French Immersion students should be given the choice of writing the tests in whatever language they prefer. Fourth, both language versions of each item can be provided next to each other on the same page in order to maximize the opportunity for the French Immersion students to understand the items.

Efficacy of the Simultaneous Test Development Approach

Simultaneous test development involves the development of tests in more than one language at the same time. Typically, bilingual item writers are recruited for test development in two languages. These item writers should have subject matter knowledge relevant to the assessment at hand as well as understanding of the characteristics of the students for whom the test is intended for. Based on the results of this study, the simultaneous approach is both efficient and effective in producing equally good tests across the two languages.

Efficiency

The simultaneous approach requires the teachers to write each item in one language and then immediately translate it into another language. They can then go back and forth to change either language version until they are convinced that both versions mean exactly the same. In this study, although four of the six item writers did not have experience in item writing and none had experience in translation, it did not take long for them to get ready for the task. The training of the teachers took no longer than three hours. These item writers also considered efficiency and speed as one attribute of the simultaneous test development approach.

It took about 18 hours for each item developer to write 30 items in both languages, close to 36 minutes per item on average. The item review and revision for the 90 items took three item writers and one researcher four hours in mathematics (about 11 minutes per item per person) and nine hours in social studies (about 24 minutes per item per person). The item review and revision for the pilot test forms (70 items) took four people four hours in each subject (about 14 minutes per item). In addition, five translators each spent about 10 hours reviewing the 58 item in mathematics and 51 items in social studies (about 28 minutes per item). The last round of revision took three research team members four hours for each subject (about 12 minutes per mathematics item and 14 minutes per social studies item). All together, it took about 101 minutes to develop one mathematics item in both languages, and 116 minutes to develop one social studies item in both languages (see Table 19).

With Alberta Education (Guimont, personal communication, November 21, 2005), it typically takes 15 to 20 minutes to write one English item in either mathematics or social studies. Then on average it takes 50 minutes to translate one mathematics item into French, and 80 minutes to translate one social studies item. Then it takes four teachers three hours to review 25 items (about 29 minutes per item). The final revision of the 25 items typically takes one translator and one developer three to five hours to complete (about 14 to 24 minutes per item). Taken together, it takes Alberta Education about 108 to 123 minutes to develop one mathematics item in both languages and 138 to 153 minutes to develop one social studies item in both languages (see Table 19). One thing worth-noting is that item translation takes about half of the total time needed to develop an item in the traditional forward translation design employed by Alberta Education. In this respect, the simultaneous approach is much more efficient in that the same teacher writes and translates each item. Given the item writers' subject knowledge

Table 19

Comparison of Item Development Time (In Minutes)

Task	Present Study		Task	Alberta Education	
	Mathematics	Social Studies		Mathematics	Social Studies
Writing	36	36	Writing	15-20	15-20
			Translation	50	80
Review 1	11	24	Review	29	29
Review 2	14	14			
Review by translators	28	28			
Revision	12	14	Revision	14-24	14-24
Total	101	116	Total	108-123	138-153

and bilingual language skills, the whole task of item writing and translation becomes less time-consuming.

To sum up, there was no indication that using the simultaneous development approach involved longer development time or higher cost than traditional test development methods, which is similar to what Solano-Flores, Trumbull, and Nelson-Barber (2002) found. Although the above calculation of time taken to write one item is a rough estimation, the simultaneous approach appears to be at least as efficient as the traditional methods. By employing bilingual teams of experts to establish the linguistic equivalence of original as well as adapted bilingual tests, the lengthy process of back translation is also bypassed.

Effectiveness

The evidence collected through the item development stage suggested that the simultaneous test development approach allowed the influence and integration of information from item writers and reviewers representing different language and cultural groups to affect test development directly. The discussions that took place extended beyond the simple choice of comparable words and phrases to the form of expressions in each language and whether differences in form would be allowed in an attempt to maintain comparable meaning while recognizing the idiomatic differences between the two languages. The item writers pointed out in particular two attributes of the simultaneous approach: reduced loss of meaning due to the immediacy of translation conducted by the same person, and deeper consideration to subtle language and culture issues in the item development process.

The item review conducted by certified translators indicated that a great majority of the items (93.1% in mathematics and 94.1% in social studies) were considered identical or similar in meaning. Following revision and selection of items, more DIF was identified in social studies (42.5%) than mathematics (19.2%), which was consistent with what was found with Alberta Grade 9 provincial tests (Gierl et al., 1999). In general, social studies tests involve more vocabulary than mathematics tests, and thus tend to produce more DIF as well. The analysis of the student protocols found no evidence for adaptation as a source of DIF in either mathematics or social studies. French Immersion students' inadequate proficiency in French appears to have contributed to the exhibition of DIF.

In tests adapted using either forward or back translation methods for English- and French-speaking examinees in Canada, adaptation differences have been identified as a source of DIF to a varying degree. Gierl, Rogers and Klinger (1999) examined adaptation-related DIF on Grade 6 provincial achievement tests, and found that 28.6% of the DIF items in mathematics were attributed to adaptation-related differences, while 26.9% of the DIF items in social studies were identified to be related to adaptation. Ercikan (2002), for example, reported success rates of 27.3% and 36.5% on TIMSS mathematics and science, respectively. Ercikan (1998) found that 44.4% of the DIF items in an international science test were linked to adaptation-related differences. Further, Ercikan et al. (2004), more recently, reported a range of 36.2% to 100% DIF items across two age groups and three content areas on the SAIP tests that were attributed to adaptation-related differences. One thing in common with all the above studies is that they all used bilingual test experts to identify possible adaptation differences, which made it impossible to compare their results and the findings of the present study directly.

Ercikan et al. (2004) conducted the only study employing think-aloud protocols for DIF analysis on adapted tests. Out of the 20 DIF items for which judgmental review identified adaptation differences as the source of DIF, six hypotheses were confirmed by students' think-aloud protocols. Although the consistency between the judgmental review and think-aloud analysis was not high, the DIF on at least five of the six items could be attributed to adaptation.

Compared with the degree to which adaptation differences contributed to DIF in the above studies, adaptation was not found to have caused DIF in any of the DIF items in mathematics or social studies. Combined with evidence from the item development process and item review, there is good reason to believe that the simultaneous approach is effective in reducing adaptation differences and producing equally good tests in two languages. Especially in the case of developing tests for second language speakers, such as French Immersion students in this study, using bilingual teachers rather than professional translators might help ensure that the language used in the tests approximate the proficiency level of the students for whom the tests are intended for.

Implications for Practice

In addition to the item writers' subject matter knowledge relevant to the assessments at hand and their experience as bilingual teachers, there are three factors that can be identified as critical to the successful implementation of the simultaneous approach. One is training and experience in item writing, which is crucial to any kind of test development. To ensure the quality of the test items in simultaneous development, it is important to provide adequate training to the bilingual teachers. More experience would also help increase the survival rate of the test items following pilot testing and item analysis. The second factor is adequate language proficiency in both languages and translation experience. The item writers should feel comfortable writing the items in both languages. Relevant translation experience would help capture the meaning through well thought-out choice of words. The third critical factor, as Solano-Flores et al. (2002) suggested, is diversity among the item writers, especially in terms of cultural backgrounds. In the present study, for example, native English speakers and native French speakers complemented each other in their item writing in both languages.

Further, the simultaneous approach is easiest to implement when only two languages are involved. When the research involves more than two languages/cultures, the task of recruiting item writers indigenous to all the languages and cultures may be a difficult challenge. The quality of the end product, however, would justify the effort made to meet the challenge.

Limitations and Directions for Future Research

The generalizability of the research was limited by the nature of the sample involved. Approximately 90% of the French-speaking examinees who participated in the field tests were bilingual French Immersion students who learned French as a second language. The interpretation of differences in performance between the English and French groups may have been confounded with the possible differences in the levels of language proficiency between the two groups. As indicated by prior research, by Grade 9 the French Immersions students would not have acquired native-like proficiency in French. While the bilingual students in this study may be representative of certain population in bilingual testing, such as English Immersion students in Hong Kong, Spanish Immersion students in the United States, and English Immersion students in South Africa, future research needs to be conducted with monolingual speakers of English and French, or other languages.

Second, Francophone students, who account for 10% of the Grade 9 French-speaking population, were not included in the interview sample of this study for practical reasons. For future research, it would be interesting to see how Francophone students interpret the test items in French. It is expected that they would not encounter the same language-related difficulty as the French Immersion students, given that French is their native language. If so, the equivalence of the test items in English and French would be further validated. However, if they share some of the difficulties with the French Immersion students on the DIF items, such as lack of familiarity with certain words or concepts, the adaptation part of the test development could be problematic.

Third, sample size posed another limitation of the current study. To keep the interview task manageable, eight students were interviewed for each item. The limitations in the sample we had might have restricted the kind of responses we got from them. In replication of the study, the size of verbal report sample, both English- and French-speaking, should be increased.

Last, one of the interviewers for the think-aloud procedure, a teacher and tutor, was found to have the tendency to prompt students at times when they were stuck solving the problems. For the purpose of this study, the researcher was interested in the way students understood the questions, not the way they solved the problems. Therefore, her interference in the interview data was partialled out to the best the researcher could. For future research, however, it is important to ensure that the interviewers follow the procedures strictly and consistently. Following adequate training, it is still necessary to monitor the way they conduct the interviews, in order to prevent problems that were not perceived or encountered in the training session.

To summarize, the simultaneous test development approach is a promising procedure with its own characteristics and advantages. More research will certainly benefit this new approach to test adaptation, especially within contexts where different languages, different content areas, and different types of assessment are involved.

Bibliography

Alberta Education. (1990). Provincial assessment of students in French immersion programs: Special report. Edmonton, AB: Student Evaluation Branch.

Alberta Education. (1991). *Language of testing study report.* Edmonton, AB: Student Evaluation Branch.

Alberta Education. (1992). *Language of testing study report.* Edmonton, AB: Student Evaluation Branch.

Alberta Learning. (2002a). Handbook for French immersion administrators. Edmonton, AB: French Language Services Branch.

Alberta Learning. (2002b). Yes, you can help! Information and inspiration for French immersion parents: Edmonton, AB: French Language Services.

Alberta Education. (2004). Grade 9 social studies achievement test school authority report: 2003-2004 school year. Edmonton, AB: Learner Assessment Branch.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the sources of differential item functioning in translated verbal items. Journal of educational *measurement*, *36*, 185-198.

Allen, M. (2004). Reading achievement of students in French immersion programs. *Educational Quarterly Review, 9*(4), 25-30.

Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Academic and the Scholastic Aptitude Rest (Report 88-2). New York, NY: College Entrance Examination Board.

Behling, O., & Law, K. S. (2000). Translating questionnaires and other research instruments: Problems and solutions. Thousand Oaks, CA: Sage.

Bentler, P., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588 –606.

Bloom, B. S. (Ed.). (1984). Taxonomy of educational objectives, Handbook 1: Cognitive domain. New York: Longman.

Bollen, K. A., & Long, J. S. (1993). Testing structural equation models. Newbury Park, CA: Sage Publications.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23*, 67-95.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-cultural Research, 1*, 185-216.

Brislin, R. W. (1973). Questionnaire wording and translation. In R. W. Brislin, W. j. Lonner, & R. M. Thorndike (Eds.), *Cross-cultural research methods* (pp. 32-58). New York: John Wiley.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-162). Newbury Park, CA: Sage.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296-322.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19*, 309-321.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Canadian Education Association. (1992). French immersion today. *CEA Information Note*, August 1992.

Carey, S. (1987). Reading comprehension in first and second languages of Immersion and Francophone students. *Canadian Journal of Exceptional Children*, 3, 103-108.

Cattell, R. B. (1956). Validation and intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology, 12*, 205-214.

Cattell, R. B., & Burdsal, C. A., Jr. (1975). The radical parcel double factoring design: A solution to the item-vs.-parcel controversy. Multivariate Behavioral Research, 10, 165-179.

CTB/McGraw-Hill. (1991). PARDUX [Computer software]. Monterey, CA: Author.

Cummins, J. (1987). Immersion programs: Current issues and future directions. In L. Stewin, & S. McCann (Eds.). *Contemporary educational issues—The Canadian mosaic* (pp. 192-206). Toronto: Copp Clark Pitman.

Day, E. M., & Shapson, S. M. (1996). *Studies in immersion education*. Clevedon, UK:

Multilingual Matters.

Dorans, N. J., & Lawrence, I. M. (1987). *The internal construct validity of the SAT* (Research Report 87-35). Princeton, NJ: educational testing Service.

Edwards, H. (1989). Review of the literature. In N. Halsall, *Immersion/regular program study*. Nepean, ON: Carleton Board of Education.

Elliott, P. R. (1994, April). *An overview of current practice in structural equation modeling*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology*, *74*, 912-920.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*, 543-553.

Ercikan, K. (April, 1999). *Translation DIF on TIMMS*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Ercikan, K. (2002). Disentangling Sources of Differential Item Functioning in Multi-language Assessments. *International Journal of Testing, 2*, 199-215.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301-321.

Ercikan, K., Law, D., Arim, R., Domene J., Lacroix, S., & Gagnon, F. (2004). *Identifying sources of DIF using think-aloud protocols: Comparing thought processes of examinees taking tests in English versus in French*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Ercikan. K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*(1), 23-35.

Erkut, S., Alarcon, O., Garcia Coll, C., Tropp, L. R., & Vaszquez Garcia, H. A. (1999). The dual-focus approach to creating bilingual measures. *Journal of Cross-Cultural Psychology, 30,* 206-218.

Ericsson, K., & Simon, H. (1993). *Protocol analysis: verbal report data.* Cambridge, MA: MIT Press.

Fraser, C. (1988). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: The University of New England.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304-312.

Genesee, F. (1987). *Learning through two languages: studies of immersion and bilingual education.* Rowley, MA: Newbury.

Gierl, M. J. (1997). Comparing the cognitive representatives of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91*, 26-32.

Gierl, M. J. (2000). Construct equivalence of translated achievement tests. *Canadian Journal of Education, 25*, 280- 296.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000). *Performance of Mantel-Haenszel, SIBTEST, and Logistic Regression when the number of DIF items is large.* Paper presented at the annual meeting of the American Educational Research Association. New Orleans.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*(2), 164-187.

Gierl, M. J., & Mulvenon, S. (1995, April). *Evaluating the application of fit indices to structural equation models in educational research: A review of the literature from 1990 though 1994.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and judgment reviews to identify and interpret differential item functioning. *Alberta Journal of Educational Research, XLV* (4), 353-376.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Greenfield, P. M. (1997). Your can't take it with you: Why ability assessments don't cross cultures. American Psychologist, 52, 1115-1124.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-cultural studies. *European Journal of Psychological Assessment*, *9*, 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*, 229-244.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. Hambleton, P. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Erlbaum.

Hambleton, R. K. & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Testing Commission*, *18*, 3-32.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Erlbaum

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45*, 153-171.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287-302.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*(3), 211-235.

Hamilton, L. S., & Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*(2), 181-200.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Holobow, N., Genesee, F., Lambert, W., & Chartrand, L. (1987). *Longitudinal evaluation of three elementary school alternatives for learning through a second language.* Montreal: Mcgill University, Department of Psychology.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Jiang, H., & Stout, W. (1998). Improved type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioural Statistics, 23*, 291- 322.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software.

Katz. I. R., Friedman, D. E., Bennett, R. E., & Berger, A. E. (1996). *Differences in strategies used to solve stem-equivalent constructed-response and multiple-choice SAT-Mathematics items* (College Board Report No. 95-03/ETS RR No. 96-20). Princeton, NJ: Educational Testing Service, and New York: the College Board.

Leighton, J. P. (2005). Avoiding Misconceptions, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice,* Winter, 6-15.

Leighton, J. P., Rogers, W. T., & Maguire, T. O. (1999). Assessment of student problem solving on ill-defined tasks. *Alberta Journal of Educational Research*, *XLV* (4), 408-426.

Lietz, P. H., & Roche, L. A. (1996, April). *Testing the invariance of reading literacy dimensions across different countries: An application of multi-group CFA.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group measurement on achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs*, No. 15.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Non-centrality and goodness of fit. *Psychological Bulletin*, *107*, 247-255.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Meara, K., & Sireci, S. G. (2003). *Appraising the dimensionality of the Medical College Admission Test.* Report submitted to the Association of American Medical Colleges, MCAT Division.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*, 107-122.

Morrison, F., and Pawley, C. (1983). *Subjects Taught in French.* Tenth Annual Report to the Ministry of Education, Part 1. Ottawa: Ottawa Board of Education Research Centre.

Nandakumar, R. (1994). Assessing latent trait unidimensionality of a set of items—Comparison of different approaches. *Journal of Educational Measurement, 31*, 1-18.

Nelson, L. R. (2000). *Item analysis for tests and surveys using LERTAP 5*. Perth, Western Australia: Curtin University of Technology.

Ramsey, P.A. (1993). Sensitivity review: The ETS experience as a case study. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning* (p. 367-389). Hillsdale, NJ: Lawrence Earlbaum.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Rogers, W. T. (2001). *Test theory*. Unpublished manuscript, University of Alberta at Alberta, Canada

Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. (2003). Differential validity and utility of successive and simultaneous approaches to the development of equivalent achievement tests in French and English. *The Alberta Journal of Educational Research, 49,* 290-304.

Romney, J., Romney, D., and Braun, C. (1989). The effects of reading aloud in French to immersion children on second language acquisition. *The Canadian Modern Language Review*, 45(3), 530-538.

Romney, J., Romney, D., and Menzies, H. (1995). Reading for pleasure in French: A study of the reading habits and interests of French immersion children. *The Canadian Modern Language Review*, 51(3), 474-509.

Roussos, L., & Stout, W. (1996a). A Multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenzel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Samuel, M. J. (1990). *Language of testing effects for academic achievement of French Immersion students*. Unpublished master's thesis, University of Alberta, Canada.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Education Statistics, 6*, 317-375.

Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16*, 12-19.

Sireci, S. G., & Berberoğlu, G. (2000). Using bilinguals to evaluate translated assessment questions. *Applied Measurement in Education, 13* (3), 229-248.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). *Adapting credentialing examinations in international uses.* Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Solano-Flores, G., Jovanovic, J., & Shavelson, R. J., & Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21*(3), 293–315.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2,* 107-129

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271-295.

Sperber, A. D., Devellis, R. F., & Boehlecke, B. (1994). Cross-cultural translation methodology and validation. *Journal of Cross-cultural Psychology*. 25(4), 501-524.

Stansfield, C. W., & Kahl, S. R. (1998). *Lessons learned from a tryout of Spanish and English versions of a state assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Streiner. D. L., & Norman, G. R. (1995). *Health measurement scales* (2nd ed.) Oxford, UK: Oxford University Press.

Stout, W., & Roussos, L. (1995). *SIBTEST Manual*. University of Illinois: Department of Statistics, Statistical Laboratory for Education and Psychological Measurement.

Stout, W., & Roussos, L. (1999). *Dimensionality-based DIF/DBF package* [Computer program].William Stout Institute for Measurement: University of Illinois.

Streiner, D. L., & Norman, G. R. (1995). Health measurement scale (2nd ed.). Oxford, UK: Oxford University Press.

Swain, M. (1974). Child bilingual language learning and linguistic interdependence. In S. Caret (Ed.), *Bilingualism, biculturalism and education* (pp. 75-81). Proceedings from the Conference at College Universitaire Saint-Jean, University of Alberta.

Swain, M., & Lapkin, S. (1981). *Bilingual education in Ontario: A decade of research.* Toronto: Ontario Institute for Studies in Education.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Tanaka, J. S. (1993). Multifaceted concepts of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA:Sage.

Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. Hambleton, P. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Erlbaum.

Triandis, H. (1976). Approaches toward minimizing translation. In R. Breslin (Ed.), *Translation: Applications and research* (pp. 229-243). New York: Wiley/Halstead.

van de Vijer, F., & Leung, K. (1997). Methods and data-analysis for cross-cultural research. Thousand Oaks, CA: Sage.

van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An Overview. *European Review of Applied Psychology, 47,* 263-279.

Werner, O., & Campbell, D. T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), *A handbook of method in cultural anthropology* (pp. 398-420). New York: The Natural History Press.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*, 136-147.

Appendix A

Grade 9 Mathematics Subject Blueprint

Reporting Category	Descriptions	Numbers	Patterns & Relations	Shape & Space	Statistics & Probability
Knowledge	• Recall facts, concepts, and terminology • Know procedures for algorithms and computations, and for using formulas • Know procedures for constructions, conversions, and order of operations • Know mental computation and estimation strategies • Know how to use calculators and computers	4	4	5	3
Skills	• Apply basic mathematical concepts in familiar and unfamiliar situations • Demonstrate relationships among number systems, operations, number forms, and concrete, pictorial, and symbolic representations • Demonstrate and apply relationships within equations and formulas • Demonstrate and apply relationships among geometric forms in a variety of situations • Demonstrate relationships between numbers and geometric forms • Use a variety of strategies to solve problems • Apply data management skills to solve problems • Judge the reasonableness of a solution	9	11	9	5
	Number (Percentage) of Questions	13 (26%)	15 (30%)	14 (28%)	8 (16%)

Note: From "Grade 9 Mathematics Subject Bulletin", by Alberta Education, retrieved October 5, 2005 from http://www.education.gov.ab.ca/k_12/testing/achievement/bulletins/Gr9_Math/gr9_math_blueprint.asp

Appendix B

Grade 9 Social Studies Subject Blueprint

		Technology and Change	Economic Systems	Quality of Life in Different Economic Systems	The Former U.S.S.R.	
Reporting Category		Industrialization, Technology	Market Economy, Mixed Economy, Centrally Planned Economy	Quality of Life	Geography, Economic Change	Number (Percentage) of Questions
Knowledge	Understands Generalizations, Concepts, Related Concepts, Terms, and Facts	9	9	2	2	22 (40%)
Skills	Locating, Interpreting, Organizing	12	12	6	3	33 (6%)
	Analyzing, Synthesizing, Evaluating					
	Number (Percentage) of Questions	21 (38%)	21 (38%)	8 (15%)	5 (9%)	55 (10%)

Note: From "Grade 9 Social Studies Subject Bulletin", by Alberta Education, retrieved October 5, 2005 from http://www.education.gov.ab.ca/k_12/testing/achievement/bulletins/Gr9_Social/gr9_soc_gen.asp#blueprint

Appendix C

Mathematics Translator Review Results

	Reviewer							
Item	1	2	3	4	5	6	Median	Range
1	3	3	3	3	3	3	3	1
2	3	3	3	3	3	3	3	1
3	1	3	3	3	3	3	3	3
4	3	3	3	3	3	3	3	1
5	3	1	3	3	3	3	3	3
6	3	3	2	3	2	3	3	2
7	3	3	3	3	2	3	3	2
8	3	3	3	3	3	3	3	1
9	3	3	3	3	3	3	3	1
10	3	3	3	3	3	3	3	1
11	3	3	3	3	3	3	3	1
12	3	2	3	3	2	3	3	2
13	3	3	3	3	3	3	3	1
14	2	2	3	3	3	3	3	2
15	3	3	3	3	3	3	3	1
16	3	3	3	3	3	3	3	1
17	3	1	2	2	2	2	2	3
18	1	1	3	1	1	3	1	3
19	3	3	3	3	3	3	3	1
20	3	3	3	3	3	3	3	1
21	3	3	3	3	3	2	3	2
22	3	3	3	3	2	3	3	2
23	3	3	3	3	3	3	3	1
24	3	3	3	3	3	3	3	1
25	3	3	3	3	3	3	3	1
26	3	3	3	3	2	3	3	2
27	3	3	3	3	1	2	3	3
28	3	3	3	1	1	1	2	3
29	1	1	1	1	3	1	1	3
30	3	3	3	3	3	3	3	1
31	3	3	3	3	3	3	3	1
32	3	1	3	3	2	3	3	3
33	3	3	3	3	3	3	3	1
34	3	3	3	3	2	3	3	2
35	3	3	3	3	3	3	3	1
36	3	3	3	3	3	3	3	1
37	3	3	3	3	3	3	3	1
38	1	1	1	1	2	1	1	2
39	3	3	3	3	3	3	3	1
40	3	3	3	3	3	3	3	1
41	3	3	3	1	3	3	3	3
42	3	2	3	3	3	3	3	2
43	3	2	3	2	3	3	3	2
44	3	3	3	3	3	3	3	1
45	3	3	3	3	3	3	3	1
46	3	3	3	3	2	3	3	2
47	3	3	3	3	3	3	3	1
48	3	3	3	3	2	2	3	2
49	3	3	3	3	3	3	3	1
50	3	3	3	3	3	3	3	1
51	3	1	3	1	1	1	1	3
52	3	3	3	3	3	2	3	2
53	3	3	3	3	2	3	3	2
54	3	2	3	3	2	2	2.5	2
55	3	3	3	3	3	2	3	2
56	3	3	3	3	2	3	3	2
57	3	3	3	3	3	3	3	1
58	3	2	2	3	3	3	3	2
JDM	7.5	11.5	7.5	4.5	17.5	8.5		

Note. Code for comparability of meaning: 1 = different, 2 = similar, and 3 = identical. Range = highest rating – lowest rating + 1.

Appendix D

Social Studies Translator Review Results

Item	Reviewer						Median	Range
	1	2	3	4	5	6		
1	3	2	2	3	2	3	2.5	2
2	3	2	3	3	3	3	3	2
3	2	1	2	1	2	2	2	2
4	2	2	3	2	2	3	2	2
5	3	3	3	3	2	3	3	2
6	3	3	2	3	2	2	2.5	2
7	2	2	2	2	2	2	2	1
8	3	3	3	3	2	1	3	3
9	3	1	3	1	1	2	1.5	3
10	3	2	2	3	2	3	2.5	2
11	3	2	3	3	2	2	2.5	2
12	3	3	3	3	2	3	3	2
13	3	2	3	3	1	3	3	3
14	3	2	3	3	2	3	3	2
15	3	3	3	3	2	2	3	2
16	3	3	3	3	2	3	3	2
17	3	1	3	3	2	3	3	3
18	1	3	3	3	3	3	3	3
19	3	2	3	3	2	2	2.5	2
20	2	1	2	3	2	3	2	3
21	3	2	3	3	2	3	3	2
22	3	2	3	3	1	3	3	3
23	3	1	1	3	1	2	1.5	3
24	3	3	3	3	2	2	3	2
25	1	1	1	2	2	2	1.5	2
26	3	3	3	3	2	1	3	3
27	3	2	3	3	2	3	3	2
28	3	3	1	3	1	2	2.5	3
29	2	3	2	3	2	3	2.5	2
30	3	1	1	2	1	3	1.5	3
31	2	3	3	2	2	3	2.5	2
32	3	1	2	2	1	2	2	3
33	3	3	2	3	2	2	2.5	2
34	2	2	2	2	2	2	2	1
35	3	2	3	3	2	3	3	2
36	2	3	3	3	2	3	3	2
37	2	3	3	3	2	1	2.5	3
38	3	3	3	3	2	3	3	2
39	2	2	2	2	2	3	2	2
40	2	3	3	3	2	3	3	2
41	1	2	3	2	1	3	2	3
42	3	3	3	3	2	-	3	2
43	1	1	2	2	1	2	1.5	2
44	1	1	1	3	2	3	1.5	3
45	2	1	1	3	1	3	1.5	3
46	3	3	3	3	2	3	3	2
47	3	2	3	2	2	3	2.5	2
48	2	2	2	3	2	3	2	2
49	3	1	3	1	1	1	1	3
50	3	3	2	3	2	2	2.5	2
51	3	2	2	2	2	2	2	2
JDM	21.5	21.5	15.5	15.5	33.5	24.5		

Note. Code for comparability of meaning: 1 = different, 2 = similar, and 3 = identical. Range = highest rating – lowest rating + 1.

Appendix E

Mathematics Achievement Test

Description	*Instructions*
• This test has 28 multiple-choice questions and 2 numerical-response questions. • This test is divided into four sections based on the major concepts studied in Grade 9 Mathematics: ➢ Number ➢ Patterns & Relations ➢ Shape & Space ➢ Statistics & Probability **This test was developed to be completed in one class period.**	• Read each question carefully and choose the **correct** or **best** answer. • Circle the letter corresponding to your answer. Example If $x = 3$, what is the value of $x + 8$? A. 10 B. 11 C. 12 D. 13 • You are expected to provide your own calculator. • Be sure that you calculator is in degree (DEG) mode. • Manipulatives may be used for this test. • Try to answer every question.

NUMBER

1. If $12^3 \times 12^{-7} \times 12^0 = 12^x$, then x equals

A. −21

B. −4

C. 0

D. $\frac{1}{12}$

2. Simplify the following expression:

$$\frac{(n^{10})^4 \div n^8}{n^2}$$

A. n^{30}

B. n^3

C. $n^{2.5}$

D. $n^{0.875}$

3. Simplify the following expression:

$$\frac{2^4 \times 2^2 \times (5^2)^3 \times 7^0}{10^3}$$

A. 7 000

B. 4 000

C. 1 000

D. 200

Use the following information to answer question 4.

> The Jacksons are going to Colorado for their summer vacation. They budgeted $3 000 Canadian to spend on their two-week trip. The rate of exchange for a Canadian dollar was 0.6302 U. S. dollars. The family's total expenses for the trip were $1 760.80 US. At the border, Mr. Jackson exchanged the American money he had left for Canadian money.

4. What is the Canadian dollar value of the Jacksons' expenses?

A. $1 109.66

B. $1 239.20

C. $1 890.60

D. $2 794.03

---

5. The students' union is selling tickets for a dance. The school is willing to contribute $250 and the tickets cost $3 per person. Which inequality represents the number of tickets that need to be sold, given that the cost of the dance is $650?

A. $\frac{650+250}{3} \geq x$

B. $250+3x \leq 650$

C. $3x \geq 650+250$

D. $400 \leq 3x$

6. If $(x^{-3})^3 = \frac{1}{512}$, then x equals

A. 8
B. 2
C. –2
D. –8

7. Car rental company A charges $32.00 plus $1.10/km while car rental company B charges $37.00 plus $0.90/km. Which company offers better price for a trip of 500 km and by how much?

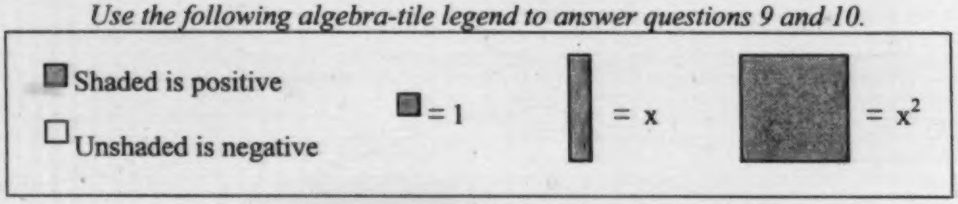
A. A by $95
B. A by $105
C. B by $95
D. B by $105

8. John can build a fence in five hours. Luke can build the same size of fence in six hours. Diane can build one in three hours. How long will the three friends take to build the fence if they all work together?

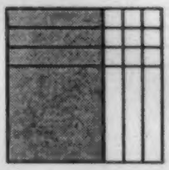
A. 5 h 07 min
B. 4 h 07 min
C. 1 h 43 min
D. 1 h 26 min

PATTERNS AND RELATIONS

Use the following algebra-tile legend to answer questions 9 and 10.


Shaded is positive
Unshaded is negative
= 1
= x
= x²

Use the following diagram to answer question 9.



9. The diagram above shows the product of which pair of polynomials?

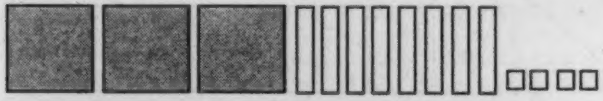
A. $(x + 3)(x + 3)$

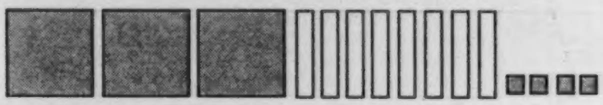
B. $(x + 3)(x - 3)$

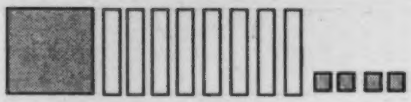
C. $(3x + 3)(3x - 3)$

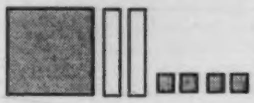
D. $(x^2 + 3x)(x^2 - 3x)$

10. Which illustration represents the difference between $(2x^2 - 5x)$ and $(-x^2 + 3x + 4)$?

A.


B.


C.


D.


11. What is the value of the expression $\frac{5x^2y-3xy}{xy}$ if $x = -2$ and $y = 4$?

A. –43

B. –37

C. –13

D. –7

12. Together, three friends have $256.00. There are three times as many $5 bills as there are $20 bills and four fewer $10 bills than $5 bills. If there are three times more loonies than $5 bills, how many $20 bills are there?

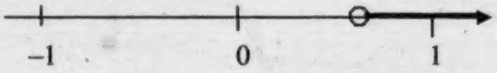
A. 4

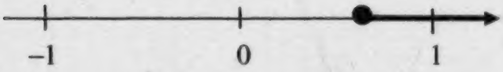
B. 8

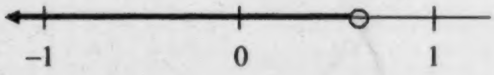
C. 12

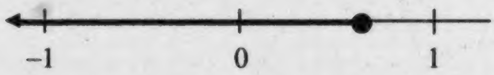
D. 16

13. Which of the following number lines represents the solution to the inequality $-2x+7>5x+3, x\in R$?

A.


B.


C.


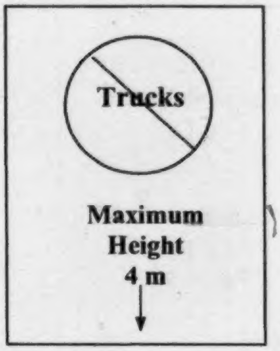
D.


14. Joe is two years older than Ben. Rita is 5 years older than the sum of the ages of Joe and Ben. The sum of the ages of the three people is 49 years. What is Rita's age?

A. 27 years
B. 29 years
C. 37 years
D. 39 years

15. Which inequality represents the maximum height of vehicles permitted in a tunnel?

A. $h > 4$ m
B. $h < 4$ m
C. $h \geq 4$ m
D. $h \leq 4$ m



16. What is the perimeter of a rectangle with a width of $(9x^2 - 5)$ and a length of $(x^2 - 3x + 3)$?

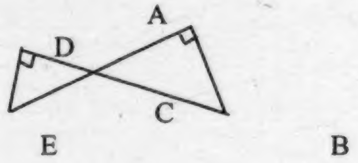
A. $10x^2 - 3x - 2$
B. $20x^2 - 6x - 4$
C. $9x^4 - 27x^3 + 22x^2 + 15x - 15$
D. $18x^4 - 54x^3 + 44x^2 + 30x - 30$

SHAPES AND SPACE

17. When point A(–5, 2) makes a 180° rotation about the origin, the coordinates of A′ will be

A. (5, –2)
B. (–5, –2)
C. (5, 2)
D. (–5, 2)

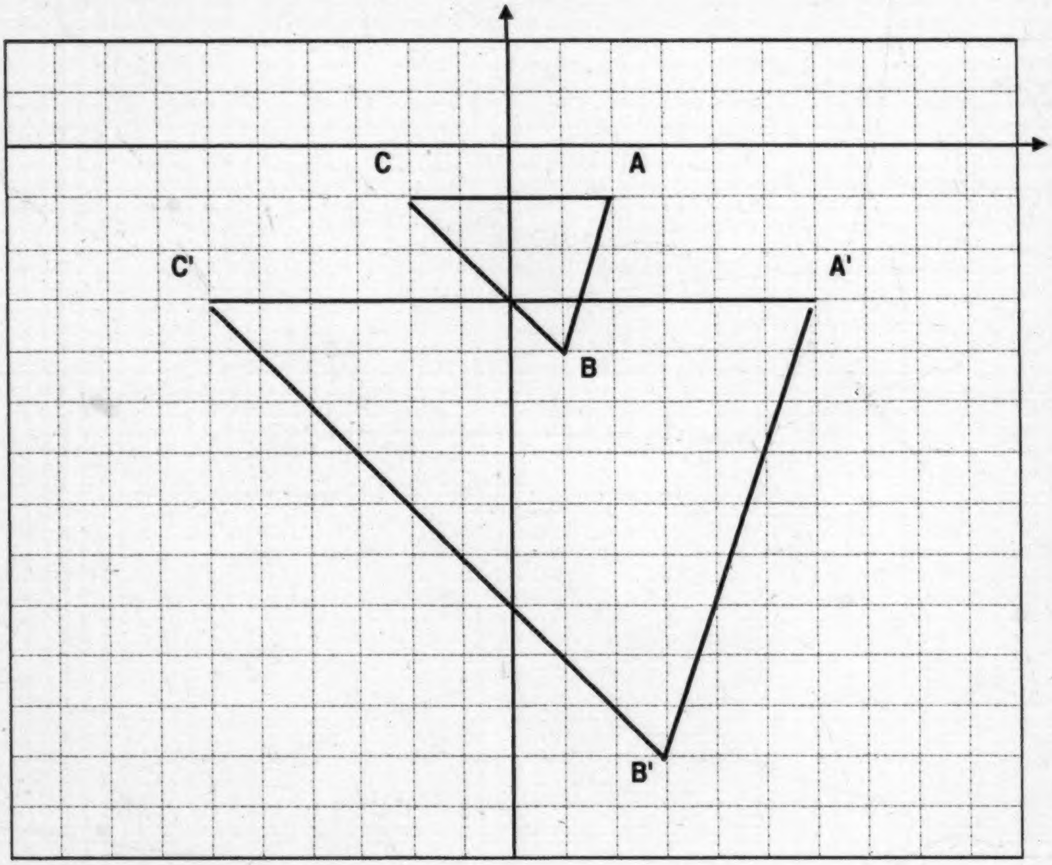
18. In the figure, $\angle ACB = 41^{\circ}$. Find $\angle DEC$.



A. 41°
B. 49°
C. 59°
D. 90°

Use the following information to answer question 19.

On the grid below, the original image is ΔABC and the dilatation image is ΔA'B'C'.



19. The scale factor of the dilatation is

A. $\frac{1}{4}$

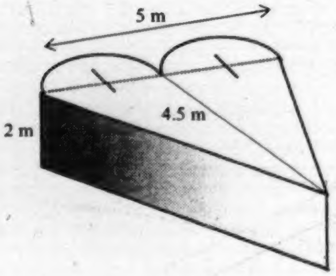
B. $\frac{1}{3}$

C. 3

D. 4

20. A heart-shaped swimming pool is built to celebrate St. Valentine's Day. Find the capacity of the pool (1 m^3 = 1 000 L).

A. 32 317 L
B. 42 135 L
C. 61 770 L
D. 179 580 L

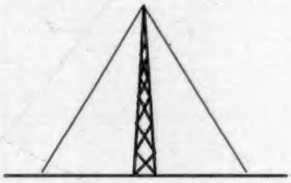


21. On the Cartesian plane, point A(–5, 1) is translated according to the rule [4, –3]. What are the coordinates of point A′ ?

A. (–9, –2)
B. (–9, 4)
C. (–1, –2)
D. (–1, 4)

22. A tower is supported by a series of cables. Each cable is 87 m long. The angle formed by the cable and the ground is 72°. What is the height of the tower?

A. 26.88 m
B. 40.84 m
C. 82.74 m
D. 267.76 m



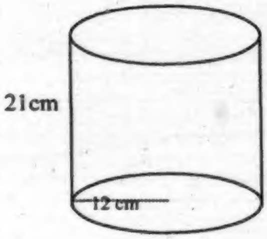
23. What is the height of a cone that has the same volume and the same base as the following cylinder?

A. 21 cm

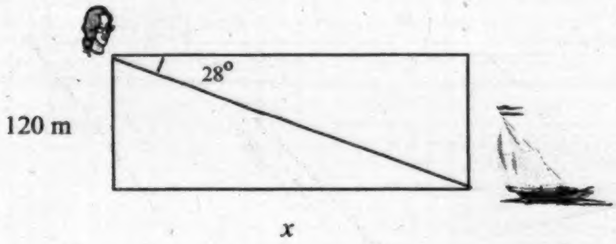
B. 63 cm

C. 162 cm

D. 9 500 cm



24. Mary Ann is observing a ship from the top of a 120 m cliff. She measures the angle of depression at 28°. What is the distance (x) from the ship to the bottom of the cliff?



A. 56 m

B. 64 m

C. 136 m

D. 226 m

STATISTICS AND PROBABILITY

25. Two regular dice are rolled. What is the theoretical probability that the sum of the two numbers on the dice is greater than 5?

A. $\frac{13}{36}$

B. $\frac{14}{36}$

C. $\frac{25}{36}$

D. $\frac{26}{36}$

26. Three regular dice are rolled. What is the probability of rolling 4, then 1, and then either 2 or 5?

A. $\frac{1}{108}$

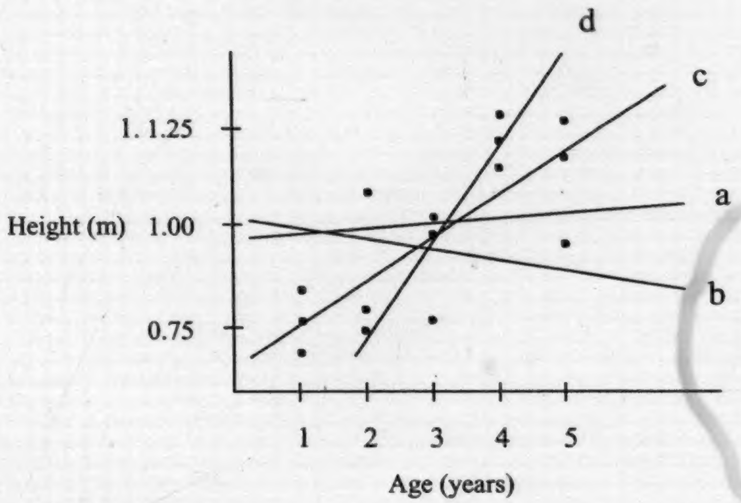
B. $\frac{2}{108}$

C. $\frac{12}{36}$

D. $\frac{24}{36}$

27. Which line is the line of best fit?

Height of Children at Different Ages



A. Line a

B. Line b

C. Line c

D. Line d

28. A bag contains six black marbles, four red marbles and eight blue marbles. What is the probability of selecting a black marble?

A. $\frac{1}{18}$

B. $\frac{1}{6}$

C. $\frac{1}{3}$

D. $\frac{1}{2}$

Appendix F

Mathématiques Test de Rendement

Description	*Directives*
• Ce test contient 28 questions à choix multiple et 2 questions à numérique. • Ce test est divisé en quatre sections basées sur les concepts fondamentaux étudiés en Mathématiques – 9^e année. ➤ Le numbre ➤ Les régularités et les relations ➤ La forme et l'espace ➤ La statistique et la probabilité **Ce test est conçu pour être complété dans une période de classe.**	• Lis attentivement chaque question et choisis la **bonne** ou la **meilleure** réponse. • Encercle la lettre qui correspond à ta réponse. Exemple Si $x=3$, quelle est la valeur de $x+8$? A. 10 B. 11 C. 12 D. 13 • Tu dois te servir de ta propre calculatrice scientifique. • Assure-toi que ta calculatrice est au mode degré (DEG). • Tu peux utiliser du matériel de manipulation. • Essaie de répondre à chaque question.

LE NOMBRE

1. Si $12^3 \times 12^{-7} \times 12^0 = 12^x$, alors x égale

A. −21

B. −4

C. 0

D. $\frac{1}{12}$

2. Simplifie l'expression suivante:

$$\frac{(n^{10})^4 \div n^8}{n^2}$$

A. n^{30}

B. n^3

C. $n^{2,5}$

D. $n^{0,875}$

3. Simplifie l'expression suivante:

$$\frac{2^4 \times 2^2 \times (5^2)^3 \times 7^0}{10^3}$$

A. 7 000

B. 4 000

C. 1 000

D. 200

Utilise l'information suivante pour répondre à la question 4.

Les Tremblay vont au Colorado pour leurs vacances d'été. Ils ont prévu de dépenser 3 000 $ canadiens pendant deux semaines. Le taux de change pour un dollar canadien était de 0,6302 $ américain. Les dépenses totales de la famille étaient de 1 760,80 $ américains. A la frontière, M. Tremblay a échangé l'argent américain qui lui restait pour de l'argent canadien.

4. Quelle est la valeur en dollars canadienne des dépenses des Tremblay?

   A. 1 109,66 $
   B. 1 239,20 $
   C. 1 890,60 $
   D. 2 794,03 $

---

5. L'association des étudiants vend des billets pour une danse. L'école est prête à contribuer 250 $ et les billets coûtent 3 $ par personne. Quelle inégalité représente le nombre de billets qu'on doit vendre, étant donné que le coût de la danse est de 650 $?

   A. $\frac{650+250}{3} \geq x$
   B. $250+3x \leq 650$
   C. $3x \geq 650+250$
   D. $400 \leq 3x$

6. Si $(x^{-3})^3 = \frac{1}{512}$, alors x égale

A. 8
B. 2
C. –2
D. –8

7. L'entreprise A de location de voiture demande 32,00 \$ plus 1,10 \$/km tandis que l'entreprise B demande 37,00 \$ plus 0,90 \$/km. Quelle entreprise offre le meilleur prix pour une distance de 500 km et de combien?

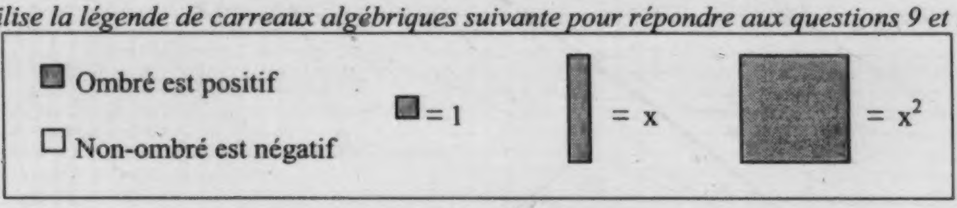
A. A de 95 \$
B. A de 105 \$
C. B de 95 \$
D. B de 105 \$

8. Jean peut bâtir une clôture en cinq heures. Luc peut bâtir une clôture de même taille en six heures. Diane peut en bâtir une en trois heures. Combien de temps les trois amis vont-ils prendre pour bâtir la clôture s'ils travaillent ensemble?

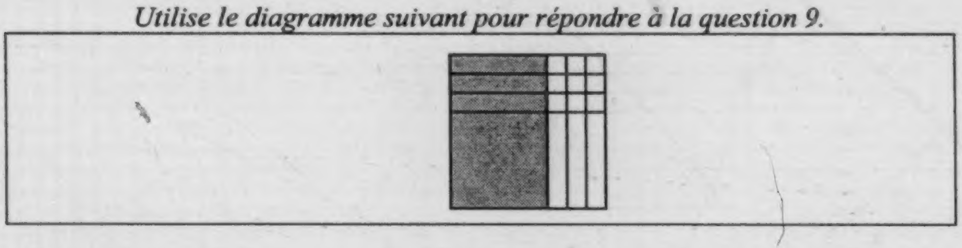
A. 5 h 07 min
B. 4 h 07 min
C. 1 h 43 min
D. 1 h 26 min

LES RÉGULARITÉS ET LES RELATIONS

Utilise la légende de carreaux algébriques suivante pour répondre aux questions 9 et 10.


Ombré est positif
Non-ombré est négatif
= 1
= x
= x²

Utilise le diagramme suivant pour répondre à la question 9.



9. Le diagramme ci-dessous montre le produit de quelle paire de polynômes?

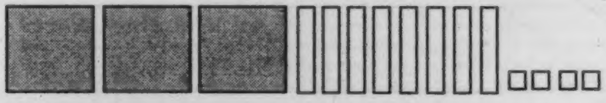
A. $(x + 3)(x + 3)$

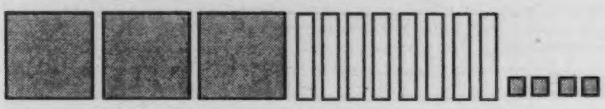
B. $(x + 3)(x - 3)$

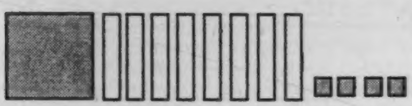
C. $(3x + 3)(3x - 3)$

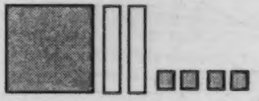
D. $(x^2 + 3x)(x^2 - 3x)$

10. Quelle illustration représente la différence entre $(2x^2 - 5x)$ et $(-x^2 + 3x + 4)$?

A.


B.


C.


D.


11. Quelle est la valeur de l'expression $\frac{5x^2y - 3xy}{xy}$ si $x = -2$ et $y = 4$?

A. –43

B. –37

C. –13

D. –7

12. Ensemble, trois amis ont 256,00 $. Il y a trois fois plus de billets de 5 $ que de billets de 20 $ et 4 billets de 10 $ de moins que de billets de 5 $. S'il y a trois fois plus de pièces de 1 $ que de billets de 5 $, combien de billets de 20 $ y a-t-il?

A. 4

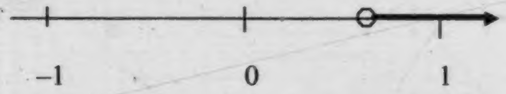
B. 8

C. 12

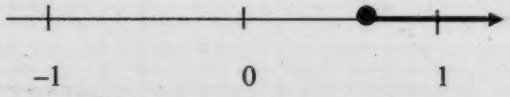
D. 16

13. Laquelle des droites numériques suivantes représente la solution à l'inégalité $-2x + 7 > 5x + 3,\ x \in R$?

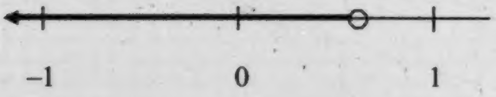
A.



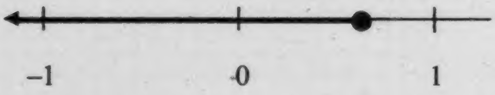
B.



C.



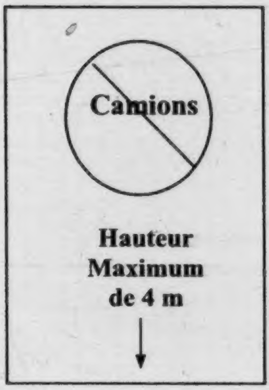
D.



14. Joe a deux ans de plus que Ben. Rita a 5 ans de plus que la somme des âges de Joe et de Ben. La somme des âges des trois personnes est de 49 ans. Quel âge a Rita?

A. 27 ans
B. 29 ans
C. 37 ans
D. 39 ans

15. Quelle inégalité représente la hauteur maximale de véhicules permises dans un tunnel?



A. $h > 4$ m
B. $h < 4$ m
C. $h \geq 4$ m
D. $h \leq 4$ m

16. Quel est le périmètre d'un rectangle avec une largeur de $(9x^2 - 5)$ et une longueur de $(x^2 - 3x + 3)$?

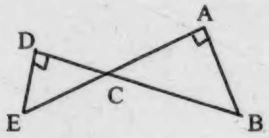
A. $10x^2 - 3x - 2$
B. $20x^2 - 6x - 4$
C. $9x^4 - 27x^3 + 22x^2 + 15x - 15$
D. $18x^4 - 54x^3 + 44x^2 + 30x - 30$

LA FORME ET L'ESPACE

17. Lorsqu'un point A(–5, 2) subit une rotation de 180° autour de l'origine, les coordonnées de A′ seront

A. (5, –2)
B. (–5, –2)
C. (5, 2)
D. (–5, 2)

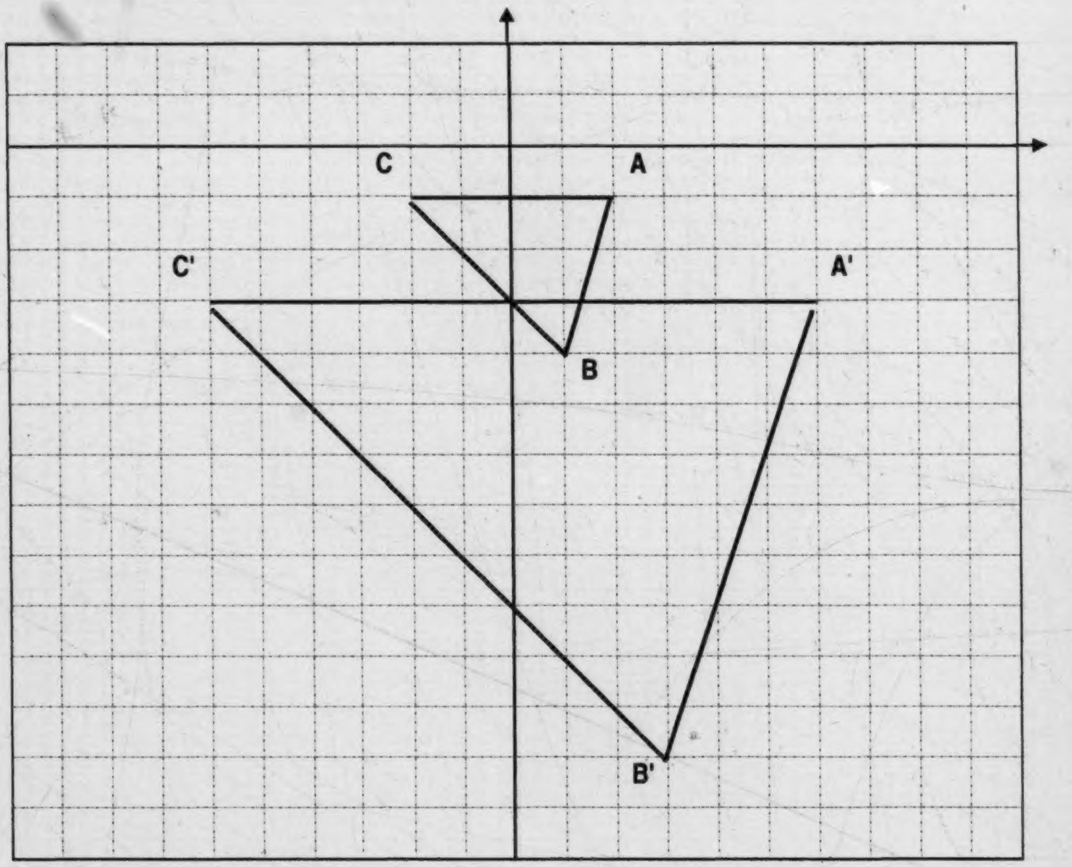
18. Dans la figure, $\angle ACB = 41°$. Trouve $\angle DEC$.



A. 41^{o}
B. 49^{o}
C. 59^{o}
D. 90^{o}

Utilise l'information suivante pour répondre à la question 19.

Sur la grille ci-dessous, l'image originale est ΔABC et l'image agrandie est ΔA'B'C'.



19. Le facteur d'échelle de l'agrandissement est

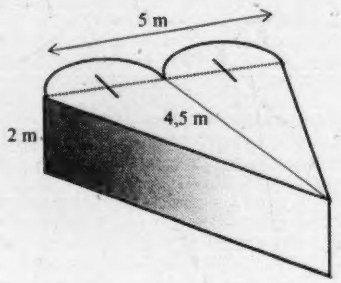
A. $\frac{1}{4}$

B. $\frac{1}{3}$

C. 3

D. 4

20. On construit une piscine en forme de coeur pour célébrer la Saint-Valentin. Trouve la capacité de la piscine (1 m^3 = 1 000 L).



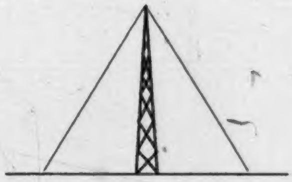
A. 32 317 L
B. 42 135 L
C. 61 770 L
D. 179 580 L

21. Sur le plan cartésien, le point A(−5, 1) subit une translation selon la règle [4, −3]. Quelles sont les coordonnées du point A′ ?

A. (−9, −2)
B. (−9, 4)
C. (−1, −2)
D. (−1, 4)

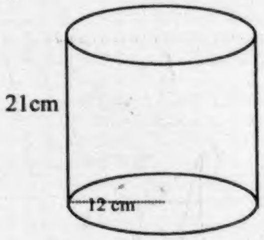
22. Une tour est supportée par une série de câbles. Chaque câble est de 87 m en longueur. L'angle formé par le câble avec la terre est de 72°. Quelle est la hauteur de la tour?

A. 26,88 m
B. 40,84 m
C. 82,74 m
D. 267,76 m

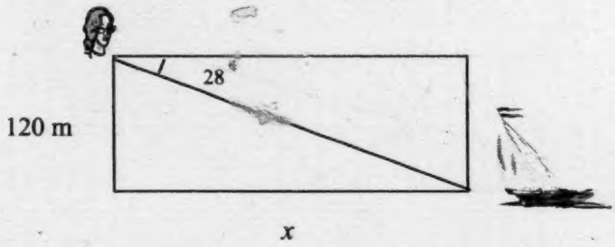


23. Quelle est la hauteur d'un cône qui a le même volume et la même base que le cylindre suivant?

A. 21 cm
B. 63 cm
C. 162 cm
D. 9 500 cm



24. Marianne observe un navire du haut d'une falaise de 120 m. Elle mesure l'angle de dépression à 28°. Quelle est la distance (x) entre le navire et le bas de la falaise?



A. 56 m
B. 64 m
C. 136 m
D. 226 m

LA STATISTIQUE ET LA PROBABILITÉ

25. On roule deux dés réguliers. Quelle est la probabilité théorique que la somme des deux nombres sur les dés soit plus grande que 5?

A. $\frac{13}{36}$

B. $\frac{14}{36}$

C. $\frac{25}{36}$

D. $\frac{26}{36}$

26. On roule trois dés réguliers. Quelle est la probabilité de rouler 4, puis 1, et ensuite soit 2 ou 5 ?

A. $\frac{1}{108}$

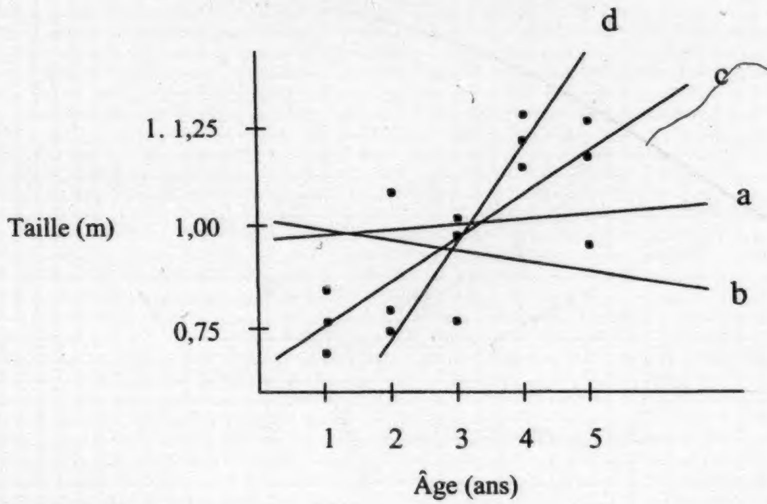
B. $\frac{2}{108}$

C. $\frac{12}{36}$

D. $\frac{24}{36}$

27. Quelle droite est la droite la mieux ajustée?

La taille des enfants à différents âges



A. la droite a

B. la droite b

C. la droite c

D. la droite d

28. Dans un sac il y a six billes noires, quatre billes rouges et huit billes bleues. Quelle est la probabilité de choisir une bille noire?

A. $\frac{1}{18}$

B. $\frac{1}{6}$

C. $\frac{1}{3}$

D. $\frac{1}{2}$

Appendix G

Social Studies Achievement Test

Description	*Instructions*
• This examination has 40 multiple-choice questions. • This test is divided into sections based on the major concepts studied in Grade 9 Social Studies: ➢ Technology and Change ➢ Economic Systems ➢ Quality of Life ➢ The Former USSR **This test was developed to be completed in one class period.**	• Read each question carefully and choose the **correct** or **best** answer. • Circle the letter corresponding to your answer. Example Edmonton is the capital city of A. Alberta B. Manitoba C. Saskatchewan D. British Columbia • Try to answer every question.

TECHNOLOGY AND CHANGE

1. What was the impact of the implementation of mass production on the workers?

   The specialization of work and

   A. a reduction in working hours.
   B. improved working conditions.
   C. more control over production.
   D. loss of control over the end product.

Use the following table to answer question 2.

The Evolution of Employment in Canada
Number of employees per sector

Sector	1931	1961	1986
Agriculture	1 131 845	640 786	479 190
Factories	442 681	1 404 865	2 196 745
Commerce	313 912	991 490	1 606 010

— adapted from *Canadians responding to change*

2. According to this table, which sector of employment lost the most workers from 1931 to 1986?

   A. Primary sector
   B. Secondary sector
   C. Tertiary sector
   D. Quaternary sector

Use the following information to answer questions 3 to 5.

Municipal workers in North Battleford are urging the city to strongly reject a proposal to enter into a public private partnership to build a new sewage treatment plant.

U.S. Filter Canada, which is owned by the French multinational Vivendi, approached the City of North Battleford with the proposal this spring. The city is currently discussing various ways of financing the construction of a new sewage treatment plant.

CUPE Local 287, which represents 123 municipal workers including sewer and water plant operators, outlined their concerns in a presentation to North Battleford's City Council on June 17:

> We feel it is extremely important to provide City Council with information about the dangers of public private partnerships," said local president Barb Plews. "We don't want the community to lose control of such a vital resource like water to a huge multinational corporation that is more interested in reaping profits than in providing good, clean drinking water.

— adapted and translated from *Syndicat Canadien de la Fonction publique*

3. According to this text, we can say that CUPE Local 287

A. supports the North Battleford city council.

B. supports the privatization of the sewer treatment plant.

C. opposes the privatization of the sewer treatment plant.

D. supports a partnership with public and private sectors in the issue of sewer treatment.

4. According to this text, what is the **most** important issue raised by CUPE Local 287?

U.S. Filter Canada

A. will not do a good job.

B. is a property of a French multinational.

C. is more interested in profit than water quality.

D. threatens the job security of the municipal workers.

5. According to this text, CUPE Local 287 is a

A. group of concerned citizens from North Battleford.

B. business specialized in sewage treatment.

C. non-profit organization.

D. workers' union.

---

6. What are the two **most** important factors that contributed to the industrialization in England?

The use of coal and the

A. development of railroads.

B. invention of the automobile.

C. mechanization of the industry.

D. development of the textile industry.

7. In 1992, the Alberta government withdrew its control over the sale of alcohol. Since then, private businesses are responsible for the sale of alcoholic beverages. This is an example of

A. privatization.

B. normalization.

C. nationalization.

D. universalization.

Use the following information to answer question 8.

URBAN AND RURAL POPULATION

Percent of Population

	1931	**1961**	**1996**
Urban	54	70	78
Rural (farming)	31	11	3
Rural (non-farming)	15	19	19

— adapted from *Statistics Canada*

8. According to this table, we can predict that in the future the population will likely

A. continue to grow in the cities and be lower in farming areas.

B. maintain itself in the cities and be lower in rural areas.

C. increase in rural areas and be lower in the cities.

D. increase in both the cities and the rural areas.

---

9. Which of the following revolutionary practices allowed Henry Ford to produce a good quality car at an affordable price?

A. Formation of monopolies

B. Introduction of closed shops

C. Invention of the assembly line

D. Cooperation of the trade unions

Use the following table to answer question 10.

CHANGES IN PRODUCTION IN THE USSR

Sectors	**1928**	**1950**	**1965**	**1978**	**1981**	**1988**
Pig Iron *(million tonnes)*	3.3	19.2	66	111	107 *(1980)*	110
Steel *(million tonnes)*	4.3	12.3	102	151	149	155
Diesel Locomotives	-	125.0	1 497	1 392	1 378 *(1980)*	n/a
Tractors *(thousands)*	1.3	116.7	405	576	555	585
Grain Combines *(thousands)*	-	46.3	101	113	106	n/a
Motor Vehicles *(millions)*	0.84	363.0	729	2 151	2 197	1 330 (estimated)
Televisions *(millions)*	-	0.01	4.9	7.2	8.2	9.6

--- adapted from *Back in the USSR*

10. Which Soviet policy **most** influenced the statistics contained in the table?

A. NEP

B. 5 Year Plans

C. Collectivization

D. Nationalization

ECONOMIC SYSTEMS

Use the following information to answer questions 11 and 12.

Speaker I
The most important value in our society is equality among people. We want to eliminate the differences among social classes.

Speaker II
In our system, the most important value is the reward for good work. If someone works hard and if he is competent, he will be rewarded.

11. In this scenario, Speaker I is in favour of a society essentially based on the principles of

A. anarchy.
B. socialism.
C. capitalism.
D. democracy.

12. In this scenario, Speaker II is in favour of a society essentially based on a

A. market economy.
B. mixed market economy.
C. centrally planned economy.
D. traditional economy.

---

13. An economic system that has the **greatest** emphasis on profits is the

A. mixed economic system.
B. market economic system.
C. traditional economic system.
D. centrally planned economic system.

Use the following information to answer question 14.

> The colonies provided primary resources and then shipped them to Great Britain, where these resources were transformed into finished products. These products were then sold to the colonies at a very high price.

14. This is an example of an economic system called

A. socialism.
B. capitalism.
C. colonialism.
D. mercantilism.

---

15. In which economic system do we witness the **most** fluctuations in the economic cycle?

A. Mixed
B. Market
C. Traditional
D. Centrally planned

16. In a market economy, the principle of equilibrium corresponds to the

A. progressive tax on revenue set up by the government.
B. subsidies provided by the government to the private sector.
C. level of government intervention in the management of the country.
D. price of consumer goods which would satisfy both consumers and producers.

Use the following information to answer questions 17 and 18.

In Canada, the government limited the right of some industries to advertise and promote their products. As an example, the tobacco companies lost the right to advertise cigarettes on television, on radio, and in magazines. Furthermore, a number of events such as the Montreal Formula 1 Automobile Grand Prix had to find new sponsors because the tobacco companies also lost their right to advertise on the racing cars or anywhere on the site of the event.

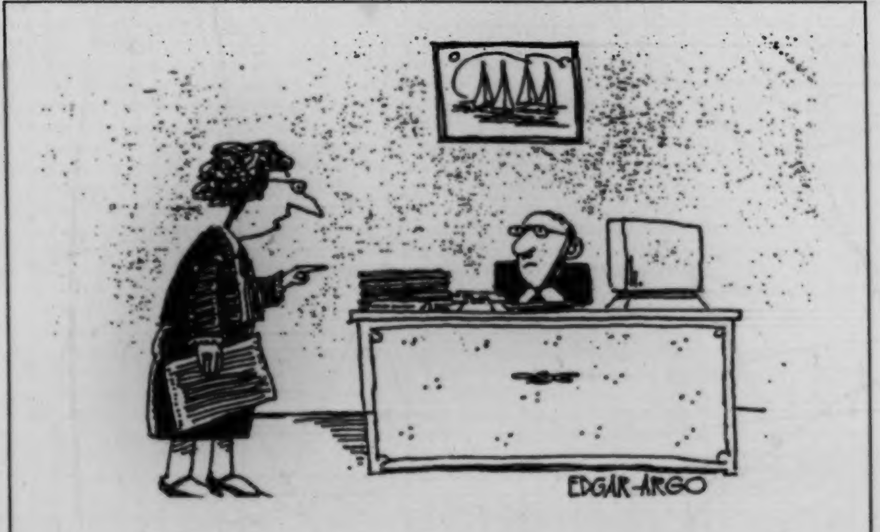
17. This action of the Canadian government is an example of

A. protectionism.
B. interventionism.
C. partnership with the private sector.
D. cooperation with the private sector.

18. If you were a shareholder in a tobacco company, what would be a major argument to oppose the government's legislation?

A. Without the right to advertise, the company will move to another country.
B. The government does not have the right to be involved in the company's business.
C. Without advertising, profits will be lower and eventually many workers will be laid off.
D. Without the right to advertise, the company will withdraw its support for events such as the Montreal Formula 1 Automobile Grand Prix.

Use the following cartoon to answer question 19.



"If all else fails, manipulate the data."

19. This situation would have **most likely** occurred in

A. Nicholas II's Russia.
B. Lenin's Russia.
C. Stalin's USSR.
D. Gorbatchev's USSR.

---

20. Which economic system would be used by a government that nationalizes some key industries like transportation, health or communication?

A. Mixed economy
B. Market economy
C. Traditional economy
D. Centrally planned economy

21. In a market economic system, which economic situation would require government intervention?

A. Bankruptcies

B. Custom duties

C. Monopolies

D. Advertising

22. One of the **main** objectives of a mixed economic system is to

A. give priority to public enterprise.

B. provide social services to consumers.

C. promote competition among private companies.

D. offer maximum subsidies to the private sector.

Use the following information to answer question 23.

Important Values of Economic Systems

1. Control + equality + group = Centrally planned
2. Intervention + collectivity + generosity = Mixed
3. Personal interest + Laissez faire + freedom = Market
4. Competition + profit + initiative = ?

23. Which economic system would complete equation 4?

A. Mixed

B. Market

C. Traditional

D. Centrally planned

Use the following information to answer questions 24 to 26.

Discussing the Kyoto Protocol, four individuals were asked to express their opinions. The Kyoto Protocol aims toward a reduction of pollution caused by fossil fuel.

Jean
I think that the government must support the Kyoto Protocol by encouraging the oil industry to experiment with other sources of energy.

Michelle
Pollution is a big problem that concerns us all. The government has the obligation to find measures to lower our gas consumption.

Paul
If the government supports the Kyoto Protocol, the oil industry will lose considerable profits. Many workers will lose their jobs.

Alexandra
I don't think that it is the government's responsibility to decrease the pollution. Each citizen has the freedom to limit or to increase his gas consumption. It's a personal choice.

24. Who would most likely support a mixed economy?

A. Jean and Paul
B. Michelle and Alexandra
C. Jean and Michelle
D. Paul and Alexandra

25. Who demonstrates the **greatest** support for a free market economy?

A. Jean
B. Michelle
C. Paul
D. Alexandra

26. Who expresses the **most** concern about the issue of unemployment?

A. Jean

B. Michelle

C. Paul

D. Alexandra

Use the following information to answer questions 27 to 29.

Swedish Woman Takes on Giants of Disposable Diapers

It wasn't money, ambition or even boredom that led Marlene Sandberg to quit her job as a corporate lawyer and become a diaper entrepreneur. It was the notion of Himalayan heaps of dirty diapers, each defying nature and refusing to decompose.

The younger of her two sons was still in diapers when Sandberg read in a newspaper about the challenge Sweden faced in getting rid of all its disposable diapers.

Concerned that her own family was contributing to the problem, she phoned around looking for companies that made biodegradable diapers. There weren't any, so she decided to try doing it herself.

Nine years later, her Nature Boy & Girl company sells diapers in some of Britain's biggest supermarkets, sharing shelf space with products from large companies such as Procter & Gamble Co. and Kimberley-Clark Corp.

Her company hopes to expand soon into France, followed by Belgium, Finland, and eventually North America.

--- adapted from *Edmonton Journal*

27. The story of Marlene Sandberg is an excellent example of

A. advertising.

B. innovation.

C. competition.

D. domestic production.

28. Marlene Sandberg's business was successful for which one of the following reasons?

A. She competed with large companies.

B. She marketed her product very well.

C. She solved the issue of pollution in Sweden.

D. She created a product more efficient than the original disposable diapers.

29. According to the law of supply and demand, what factor is the **supply** in Marlene Sandberg's story?

A. Disposable diapers

B. Biodegradable diapers

C. Children's need for diapers

D. People's concern about the environment

QUALITY OF LIFE

30. A supporter of a market economy system would say that quality of life largely depends on

A. a large array of social services offered to the public.

B. the spirit of innovation and enterprise promoted by society.

C. large subsidies available to the entrepreneurs showing initiative.

D. shared social and economic responsibility between the government and the private sector.

31. A supporter of a mixed economic system would say that the **most** important indicator of quality of life is

A. guarantee of a job for life.

B. access to social services.

C. protection of the environment.

D. choice of consumer goods.

Use the following table to answer question 32 and 33.

Immigrants from Canada to the United States by State of Residence

State	1997	1998	1999	2000
California	1 339	1 396	943	1 999
Florida	1 396	1 075	846	2 011
New York	890	755	699	1 153
Washington	657	708	529	936
Michigan	799	663	662	849
Texas	742	495	564	1 270

—adapted from *Statistical Yearbook, U.S. Immigration and Naturalization Services*

32. Which American state was the **least** successful in attracting Canadian immigrants from 1997 to 2000?

A. New York

B. Washington

C. Michigan

D. Texas

33. In which year was it **most likely** that there was an economic slow down in the United States?

A. 1997

B. 1998

C. 1999

D. 2000

Use the following information to answer questions 34 to 36.

In 1994, the government closed a large public hospital with 830 hospital beds in a large Canadian city. The hospital was then sold to private investors who reopened the hospital with 540 private hospital beds. During the samé period, the city's population increased by 16%.

We asked some residents their opinions about closing the hospital beds.

Richard
I think that the government had no other choice. Health care is very expensive and we could not afford to maintain all these hospitals.

Marie
This is terrible! I understand that maintaining a hospital is very expensive, but public health care must be a priority.

Gilbert
The government decided to close the hopital because it wants to encourage private health care. This is not acceptable!

Denise
I think that it would be normal for people who are willing to pay to have access to private health care. This will reduce the waiting time in public hospitals.

—adapted and translated from *Syndicat Canadien de la Fonction publique*

34. Who would be in favour of raising taxes to maintain public health care?

A. Richard and Marie

B. Gilbert and Denise

C. Richard and Denise

D. Marie and Gilbert

35. Who **most** supports the concept of universality of public health care system?

A. Richard and Marie
B. Gilbert and Denise
C. Richard and Denise
D. Marie and Gilbert

36. Who **most** supports the principles of a market economy?

A. Richard and Marie
B. Gilbert and Denise
C. Richard and Denise
D. Marie and Gilbert

THE FORMER USSR

Use the following information to answer question 37.

> It is not what strangers think of the Perestroïka that is important, say the Soviets, what matters is what happens here. Within the last 5 years, instead of improving, the situation has become worse. The grocery stores offer fewer products and the stores, with their poor quality and their old-fashioned clothes, look more and more like the Salvation Army store.

—adapted and translated from *La Face cachée de la Perestroïka*

37. What is the problem identified in this paragraph?

A. Scarcity
B. Repression
C. Corruption
D. The black market

Use the following information to answer question 38.

> The more things change, the more they are the same.

38. Which of the following statements best corresponds to the pattern of the history of the USSR/Russia in the 20th century, as reflected in the quotation?

A. Lenin has been a symbol of renewal for the Russian society.
B. Stalin's regime distanced itself from the autocratic regime of Imperial Russia.
C. The KGB was similar in its methods to the Secret Police of the Russian Tzars.
D. The collectivization under Stalin liberated the peasants from political oppression.

Use the following information to answer question 39.

Moscow Communists are Nostalgic

Pro-communist protesters shout anti-government slogans as they rally holding the Soviet hammer and sickle flag to mark the 82nd anniversary of the Bolshevik Revolution in downtown Moscow on Sunday. The Russian Communist Party on Saturday said that the government's rejection of socialism had resulted in "an unprecedented national humiliation."

39. The **best** conclusion that can be drawn from the newspaper article is that

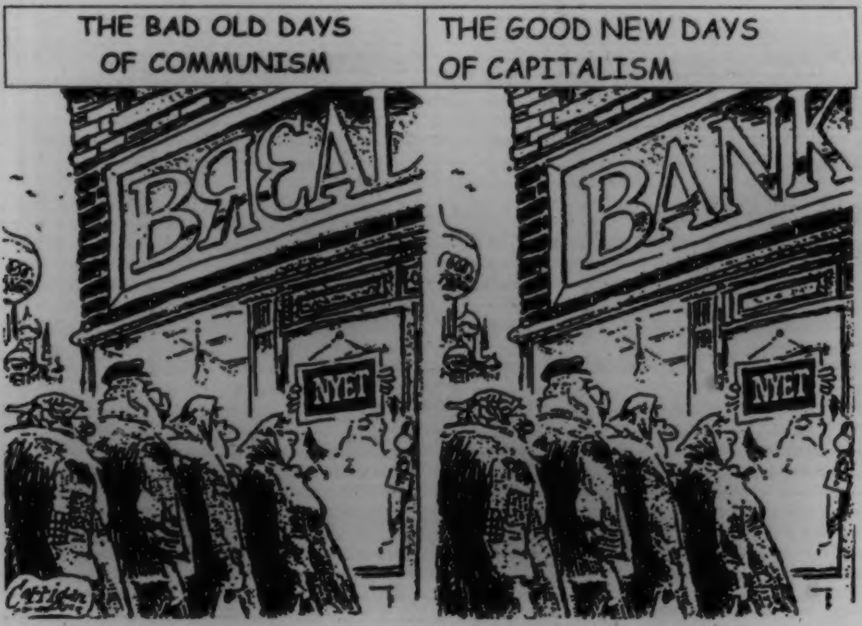
A. the public will never support the Russian government as long as it rejects communism.

B. the government's rejection of socialism created uncertainty for some Russian citizens.

C. the Russian government made a mistake when it adopted capitalism.

D. no political system can ever satisfy the public.

Use the following cartoon to answer question 40.


THE BAD OLD DAYS OF COMMUNISM
THE GOOD NEW DAYS OF CAPITALISM
BЯEA
NYET
BANK
NYET
Corrigan

40. This political cartoon **best** reflects the idea that

A. economic change in Russia has not improved quality of life.

B. Perestroïka put an end to the chronic shortage of consumer goods.

C. the bureaucracy of banks prevents a genuine transition to capitalism.

D. capitalism is the best possible solution to solve the problems of communism.

Appendix H

Études Sociales Test de Rendement

Description	*Directives*
• Ce livret contient 40 questions à choix multiple. • Ce test est divisé en quatre sections basées sur les concepts fondamentaux étudiés en Études sociales – 9^{e} année: ➢ La technologie et le changement ➢ Systèmes économiques ➢ Qualité de vie ➢ L'ancienne URSS **Ce test est conçu pour être complété dans une période de classe.**	• Lis attentivement chaque question et choisis la **bonne** ou la **meilleure** réponse. • Encercle la lettre qui correspond à ta réponse. Exemple Edmonton est la capitale A. de l'Alberta B. du Manitoba C. de la Saskatchewan D. de la Colombie-Britannique • Essaie de répondre à chaque question.

LA TECHNOLOGIE ET LE CHANGEMENT

1. Quel effet a eu l'implantation de la production en série sur les travailleurs? La spécialisation du travail et

A. la réduction des heures de travail.

B. l'amélioration des conditions de travail.

C. plus de contrôle sur la production.

D. la perte de contrôle sur le produit final.

Utilise le tableau suivant pour répondre à la question 2.

L'évolution des emplois au Canada

Nombre d'employés par secteur

Secteur	**1931**	**1961**	**1986**
Agriculture	1 131 845	640 786	479 190
Usines	442 681	1 404 865	2 196 745
Commerce	313 912	991 490	1 606 010

2. Selon ce tableau, quel secteur de travail a perdu **le plus** de travailleurs de 1931 à 1986?

A. Le secteur primaire

B. Le secteur secondaire

C. Le secteur tertiaire

D. Le secteur quaternaire

Utilise l'information suivante pour répondre aux questions 3 à 5.

Les travailleurs municipaux de North Battleford font pression sur la municipalité pour qu'elle rejette fermement un projet de partenariat public-privé pour la construction d'une nouvelle station d'épuration des eaux d'égout.

C'est U.S. Filter Canada, une société appartenant à la multinationale française Vivendi qui a, ce printemps, proposé ce projet à la ville de North Battleford. La ville étudie présentement les diverses façons de financer la construction d'une nouvelle station d'épuration des eaux d'égout.

La section locale 287 du SCFP, qui représente 123 travailleurs municipaux, dont les opérateurs des services de l'eau potable et des égouts, a fait part de ses préoccupations à la réunion du conseil de ville de North Battleford du 17 juin:

> Nous croyons qu'il est très important d'informer la ville des dangers des partenariats public-privé, a déclaré la présidente de la section locale, Barb Plews. Nous voulons à tout prix éviter que la collectivité perde le contrôle d'une ressource aussi vitale que l'eav au profit d'une énorme multinationale beaucoup plus intéressée à récolter des profits qu'à fournir une eau potable propre et de qualité à la population.

--- adapté de *Syndicat Canadien de la Fonction publique*

3. Selon ce texte, on peut affirmer que la section locale 287 du SCFP

A. appuie le conseil de ville de North Battleford.

B. appuie la privatisation de la station d'épuration des eaux d'égout.

C. s'oppose à la privatisation de la station d'épuration des eaux d'égout.

D. appuie un partenariat public-privé dans le dossier de d'épuration des eaux d'égout.

4. Selon ce texte, quelle est la préoccupation **principale** de la section locale 287 du SCFP? U.S. Filter Canada

A. ne fera pas un bon travail.

B. est la propriété d'une multinationale française.

C. est plus intéressée par les profits que par la qualité de l'eau.

D. est une menace pour les emplois des employés municipaux.

5. Selon ce texte, la section locale 287 du SCFP est

   A. un groupe de citoyens inquiets de North Battleford.
   B. une entreprise d'épuration des eaux d'égout.
   C. une association à but non-lucratif.
   D. un syndicat de travailleurs.

---

6. Quels sont les deux facteurs **les plus** importants qui ont contribué à l'industrialisation en Angleterre ?

   L'utilisation du charbon et

   A. le développement du chemin de fer.
   B. l'invention de l'automobile.
   C. la mécanisation de l'industrie.
   D. le développement de l'industrie du textile.

7. En 1992, le gouvernement de l'Alberta a cédé son contrôle de la vente d'alcool. Depuis cette date, les entreprises privées sont responsables de la vente des boissons alcoolisées. Ceci est un exemple de

   A. privatisation.
   B. normalisation.
   C. nationalisation.
   D. universalisation.

Utilise l'information suivante pour répondre à la question 8.

POPULATION URBAINE ET RURALE

Pourcentage de la population

	1931	**1961**	**1996**
Urbaine	54	70	78
Rurale (agricole)	31	11	3
Rurale (non-agricole)	15	19	19

8. Selon ce tableau, on peut prédire que, probablement, dans l'avenir la population

A. des villes va continuer à augmenter et celle des régions agricoles va diminuer.

B. des villes va se maintenir et celle des régions rurales va diminuer.

C. des régions rurales va augmenter et celle des villes va diminuer.

D. des villes et des régions rurales va augmenter.

9. Laquelle de ces pratiques révolutionnaires a permis à Henry Ford de créer une voiture de bonne qualité à un prix modéré?

A. La formation de monopoles

B. L'introduction d'ateliers fermés

C. L'invention de la chaîne de montage

D. La coopération des syndicats

Utilise le tableau suivant pour répondre aux question 10.

CHANGEMENT DE PRODUCTION EN URSS

Secteurs	**1928**	**1950**	**1965**	**1978**	**1981**	**1988**
Fonte brute *(millions de tonnes)*	3,3	19,2	66	111	107 *(1980)*	110
Acier *(millions de tonnes)*	4,3	12,3	102	151	149	155
Locomotive diesel	-	125,0	1 497	1 392	1 378 *(1980)*	s.o.
Tracteurs *(milliers)*	1,3	116,7	405	576	555	585
Moissonneuses batteuses (milliers)		46,3	101	113	106	s.o.
Véhicules à moteur (millions)	0,84	363,0	729	2 151	2 197	1330 (estimation)
Télévisions *(millions)*	-	0,01	4,9	7,2	8,2	9,6

— adapté de *Retour en URSS*

10. Quelle politique soviétique a **le plus** influencé les statistiques contenues dans ce tableau?

A. La NEP

B. Les plans quinquennaux

C. La collectivisation

D. La nationalisation

SYSTÈMES ÉCONOMIQUES

Utilise l'information suivante pour répondre aux questions 11 et 12.

Interlocuteur I
La valeur la plus importante dans notre société est l'égalité entre les personnes. Nous voulons éliminer les différences entre les classes sociales.

Interlocuteur II
Dans notre systeme, la valeur la plus importante est la récompense du travail bien fait. Si quelqu'un travaille fort et s'il est compétent, il sera récompensé.

11. Dans ce scénario, l'interlocuteur I est en faveur d'une société essentiellement basée sur des principes

A. anarchistes.
B. socialistes.
C. capitalistes.
D. démocratiques.

12. Dans ce scénario, l'interlocuteur II est en faveur d'une société essentiellement basée sur

A. une économie de marché.
B. une économie mixte.
C. une économie planifiée.
D. une économie traditionnelle.

---

13. Un système économique qui met la **plus grande** emphase sur le profit est le système d'économie

A. mixte.
B. de marché.
C. traditionnel.
D. planifié.

Utilise l'information suivante pour répondre à la question 14.

Les colonies fournissaient des ressources premières et les expédiaient en Grande-Bretagne, où elles étaient transformées en produits finis. Ces produits étaient ensuite vendus aux colonies à prix fort.

14. Ceci est un exemple de système économique appelé

A. socialisme.
B. capitalisme.
C. colonialisme.
D. mercantilisme.

---

15. Dans quel système économique voyons-nous **le plus** de fluctuations dans le cycle économique?

A. Mixte
B. De marché
C. Traditionnel
D. Planifié

16. Dans une économie de marché, le principe d'équilibre correspond

A. aux impôts progressifs sur le revenu mis en place par le gouvernement.
B. aux subventions gouvernementales fournies au secteur privé.
C. au niveau d'intervention gouvernementale dans la gestion du pays.
D. au prix des biens de consommation qui satisferait à la fois les consommateurs et les producteurs.

Utilise l'information suivante pour répondre aux questions 17 et 18.

Au Canada, le gouvernement a limité le droit de certaines industries à faire de la publicité pour faire la promotion de leurs produits. Par exemple, les compagnies de tabac ont perdu le droit de faire la promotion des cigarettes à la télévision, à la radio et dans les magazines. De plus, certains événements comme le Grand Prix Automobile Formule 1 de Montréal ont dû chercher d'autres commanditaires parce que les compagnies de tabac ont aussi perdu le droit de s'afficher sur les voitures de course ou sur le site de l'événement.

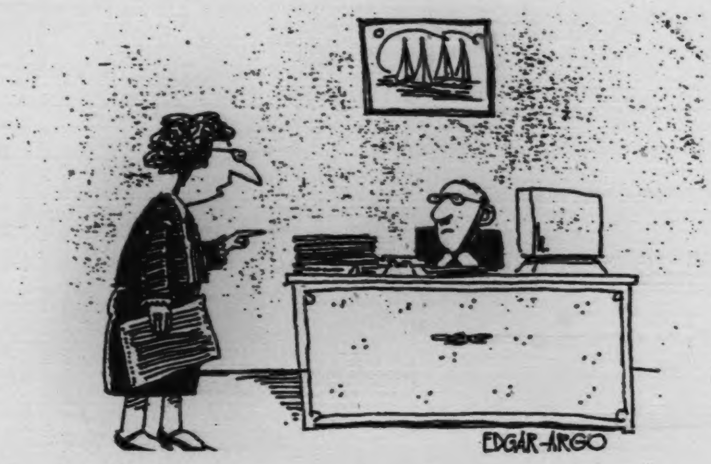
17. L'action du gouvernement canadien est un exemple

A. de protectionnisme.
B. d'interventionnisme.
C. de partenariat avec le secteur privé.
D. de coopération avec le secteur privé.

18. Si tu étais un actionnaire dans une compagnie de tabac, quel serait un argument majeur que tu utiliserais pour t'opposer à la législation du gouvernement?

A. Sans droit de publicité, la compagnie va déménager dans un autre pays.
B. Le gouvernement n'a pas le droit d'intervenir dans les affaires de la compagnie.
C. Sans publicité, les profits vont diminuer et finalement plusieurs employés vont perdre leur emploi.
D. Sans droit de publicité, la compagnie va retirer son support aux événements comme le Grand Prix Automobile Formule 1 de Montréal.

Utilise la caricature suivante pour répondre à la question 19.



« Si rien ne marche, manipule les données »

19. Cette situation se **serait surtout** produite dans

A. la Russie de Nicolas II.

B. la Russie de Lénine.

C. l'URSS de Staline.

D. l'URSS de Gorbatchev.

---

20. Quel système économique serait utilisé par un gouvernement qui nationalise certaines industries clés telles que le transport, la santé ou les communications?

A. L'économie mixte

B. L'économie de marché

C. L'économie traditionnelle

D. L'économie planifiée

21. Dans une économie de marché, quelle situation économique exigerait une intervention gouvernementale?

A. Les faillites

B. Les droits de douane

C. Les monopoles

D. La publicité

22. Un des objectifs **principaux** du système d'économie mixte est de

A. donner la priorité aux entreprises publiques.

B. offrir des services sociaux aux consommateurs.

C. promouvoir la concurrence entre les entreprises privées.

D. offrir des subventions maximales au secteur privé.

Utilise l'information suivante pour répondre à la question 23.

Valeurs importantes des systèmes économiques

1. Contrôle + égalité + groupe = Économie planifiée
2. Interventionnisme + collectivité + générosité = Économie mixte
3. Intérêt personnel + laissez-faire + liberté = Économie de marché
4. Concurrence + profit + initiative = ?

23. Quel système économique compléterait l'équation 4?

A. Mixte

B. De marché

C. Traditionnel

D. Planifié

Utilise l'information suivante pour répondre aux questions 24 à 26 .

Dans une discussion sur l'accord de Kyoto, on a demandé à quatre individus d'exprimer leur opinion. L'accord de Kyoto vise la réduction de la pollution dûe aux combustibles fossiles.

Jean
Je pense que le gouvernement doit appuyer l'accord de Kyoto en encourageant l'industrie pétrolière à expérimenter d'autres formes d'énergie.

Michelle
La pollution est un gros problème qui nous concerne tous. Le gouvernement a l'obligation d'intervenir pour diminuer notre consommation d'essence.

Paul
Si le gouvernement appuie l'accord de Kyoto, l'industrie pétrolière va perdre trop d'argent. Plusieurs personnes vont perdre leur emploi.

Alexandra
Je ne pense pas que ce soit la responsabilité du gouvernement de diminuer la pollution. Chaque citoyen a la liberté de diminuer sa consommation d'essence ou de consommer davantage. C'est un choix personnel.

24. Qui serait probablement le plus en faveur d'un système d'économie mixte?

A. Jean et Paul
B. Michelle et Alexandra
C. Jean et Michelle
D. Paul et Alexandra

25. Qui manifeste **le plus** d'appui pour une économie de marché libre?

A. Jean
B. Michelle
C. Paul
D. Alexandra

26. Qui manifeste **le plus** d'inquiétude pour le problème du chômage?

A. Jean
B. Michelle
C. Paul
D. Alexandra

Utilise l'information suivante pour répondre aux questions 26 à 28.

Une Suédoise défie les géants des couches jetables

Ce n'était pas l'argent, l'ambition ni même l'ennui qui ont mené Marlene Sandberg à quitter son emploi d'avocate pour devenir entrepreneur d'une compagnie de couches. C'était plutôt la notion de montagnes de couches sales défiant la nature en refusant de se décomposer.

Le plus jeune de ses deux fils était toujours aux couches quand Sandberg a lu dans un journal le problème des couches jetables en Suède.

Préoccupée par le fait que sa propre famille contribuait à ce problème, elle a essayé de trouver des compagnies qui fabriquaient des couches jetables biodégradables. En réalisant qu'il n'y en avait aucune, elle a décidé d'en fabriquer elle-même.

Neuf ans plus tard, sa compagnie Nature Boy & Girl vend des couches dans certains des plus grands supermarchés d'Angleterre, partageant le marché avec des grandes compagnies telles que Procter & Gamble Co. et Kimberley-Clark Corp.

Sa compagnie espère s'étendre bientôt en France, en Belgique, en Finlande et finalement en Amérique du Nord.

— adaptés et traduite de *The Edmonton Journal*

27. L'histoire de Marlene Sandberg est un excellent exemple de

A. publicité.
B. innovation.
C. concurrence.
D. production domestique.

28. L'entreprise de Marlene Sandberg a eu du succès pour la quelle des raisons suivantes?

A. Elle faisait concurrence à de grandes compagnies.

B. Elle a bien commercialisé son produit.

C. Elle a réglé le problème de la pollution en Suède.

D. Elle a créé un produit plus efficace que les premières couches jetables.

29. Selon la loi de l'offre et de la demande, quel facteur représente **l'offre** dans l'histoire de Marlene Sandberg?

A. Les couches jetables

B. Les couches biodégradables

C. Le besoin des enfants pour des couches pour

D. La préoccupation du public pour l'environnement

QUALITÉ DE VIE

30. Un partisan de l'économie de marché dirait que la qualité de vie dépend en grande partie

A. du large éventail de services sociaux offerts au public.
B. de l'esprit d'initiative et d'entreprise encouragé par la société.
C. de larges subventions disponibles aux entrepreneurs montrant de l'initiative.
D. de la responsabilité sociale et économique partagée entre le gouvernement et le secteur privé.

31. Un partisan de l'économie mixte dirait que l'indicateur **le plus** important de la qualité de vie est

A. la garantie d'un emploi pour la vie.
B. l'accès aux services sociaux.
C. la protection de l'environnement.
D. le choix de biens de consommation.

Utilise le tableau suivant pour répondre aux questions 32 et 33.

Immigrants canadiens aux États-Unis d'après leur état de résidence

États	**1997**	**1998**	**1999**	**2000**
Californie	1 339	1 396	943	1 999
Floride	1 396	1 075	846	2 011
New York	890	755	699	1 153
Washington	657	708	529	936
Michigan	799	663	662	849
Texas	742	495	564	1 270

— adapté et traduit de *Statistical Yearbook, U.S. Immigration & Naturalization Services*

32. Quel état américain a eu **le moins** de succès à attirer des immigrants canadiens de 1997 à 2000?

A. New York

B. Washington

C. Michigan

D. Texas

33. En quelle année était-il **plus probable** d'observer un ralentissement économique aux États-Unis?

A. 1997

B. 1998

C. 1999

D. 2000

Utilise l'information suivante pour répondre aux questions 34 à 36.

En 1994, le gouvernement a fermé un grand hôpital public avec 830 lits dans une grande ville canadienne. L'hôpital a ensuite été vendu à des investisseurs privés qui ont rouvert l'hôpital avec 540 lits privés. Pendant la même période, la population de la ville a augmenté de 16%.

On a demandé à quelques résidents leur opinion sur la fermeture des lits d'hôpitaux.

Richard
Je pense que le gouvernement n'avait pas le choix. Les services de santé coûtent très cher et on ne pouvait pas entretenir tous les lits d'hôpitaux.

Marie
C'est terrible! Je comprends que maintenir un hôpital coûte cher mais les services de santé doivent être une priorité.

Gilbert
Le gouvernement a décidé de fermer l'hôpital parce qu'il veut encourager les services de santé privés. C'est inacceptable!

Denise
Je pense que c'est normal que les personnes qui veulent payer puissent avoir accès à des soins de santé privés. Ceci va diminuer les files d'attente dans les hôpitaux publics.

--- adaptés de *Syndicat Canadien de la Fonction Publique*

34. Qui seraient en faveur d'augmenter les taxes pour maintenir les services de santé publics?

A. Richard et Marie

B. Gilbert et Denise

C. Richard et Denise

D. Marie et Gilbert

35. Qui appuient **le plus** le concept d'universalité des services de santé publics?

A. Richard et Marie
B. Gilbert et Denise
C. Richard et Denise
D. Marie et Gilbert

36. Qui appuient **le plus** les principes d'une économie de marché?

A. Richard et Marie
B. Gilbert et Denise
C. Richard et Denise
D. Marie et Gilbert

L'ANCIENNE URSS

Utilise l'information suivante pour répondre à la question 37.

> Ce n'est pas ce que les étrangers pensent de la Perestroïka qui est important, disent les Soviétiques, c'est ce qui se passe ici. Depuis 5 ans, loin de s'améliorer, la situation économique du pays s'est aggravée. Les comptoirs d'alimentation offrent de moins en moins de produits et les magasins, avec leur marchandise de mauvaise qualité et leurs vêtements démodés, ressemblent de plus en plus à des comptoirs de l'Armée du Salut.

--- adaptés de *La Face cachée de la Perestroïka*

37. Quel est le problème évoqué dans ce paragraphe?

A. La pénurie
B. La répression
C. La corruption
D. Le marché noir

Utilise l'information suivante pour répondre à la question 38.

> Plus ça change, plus c'est pareil.

38. Lequel des énoncés suivants correspond le mieux au tracé de l'histoire de l'URSS/Russie au 20^{e} siècle évoqué dans la citation?

A. Lénine a été un symbole de renouveau pour la société russe.
B. Le régime de Staline s'est éloigné du régime autocratique de la Russie impériale.
C. Le KGB était semblable dans ses méthodes à la police secrète des Tsars russes.
D. La collectivisation sous Staline a libéré les paysans de l'oppression politique.

Utilise l'information suivante pour répondre à la question 39.

Les Communistes se sentent nostalgiques à Moscou

Des manifestants en faveur du communisme crient des slogans contre le gouvernement alors qu'ils manifestent en tenant le drapeau soviétique du marteau et de la faucille pour marquer le 82^{e} anniversaire de la révolution bolchévique au centre ville de Moscou dimanche. Le parti communiste russe a déclaré samedi que le rejet du socialisme par le gouvernement avait eu comme résultat « une humiliation nationale sans précédent ».

39. La **meilleure** conclusion que l'on peut tirer de l'article de journal est que

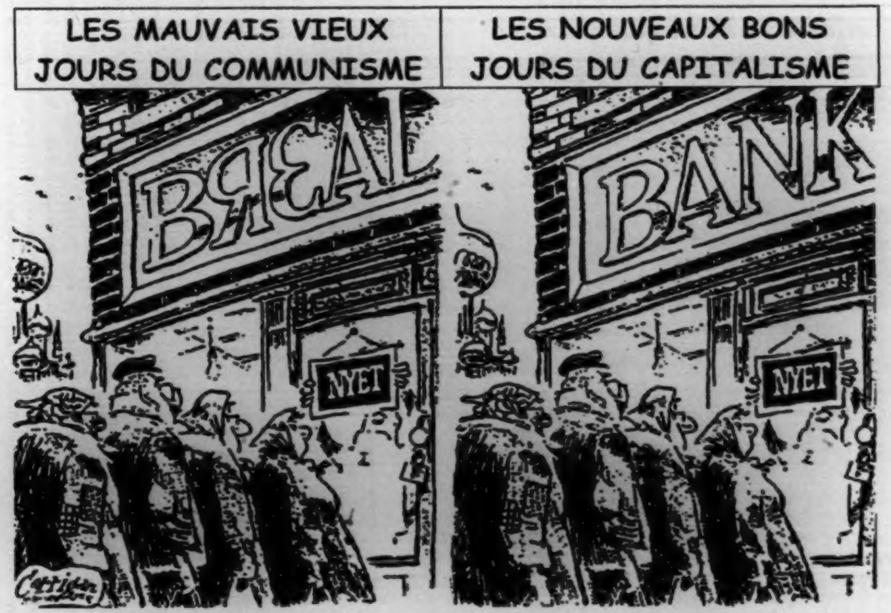
A. le gouvernement russe n'aura pas l'appui du public tant qu'il rejettera le communisme.

B. le rejet du socialism par le govemement représente une incertitude pour des citoyens russes.

C. le gouvernement russe a fait une erreur en adoptant le capitalisme.

D. aucun système politique ne peut satisfaire le public.

Utilise la caricature suivante pour répondre à la question 40.


LES MAUVAIS VIEUX JOURS DU COMMUNISME
LES NOUVEAUX BONS JOURS DU CAPITALISME
BREAD
NYET
BANK
NYET
Carrigan

40. Cette caricature de nature politique reflète **le mieux** l'idée que

A. le changement économique en Russie n'a pas amélioré la qualité de vie.

B. la Perestroïka a mis fin à la pénurie chronique de biens de consommation.

C. la bureaucratie des banques empêche une véritable transition au capitalisme.

D. le capitalisme est la meilleure solution possible pour résoudre les problèmes du communisme.

Appendix I

Parent Consent Letter

May 10, 2005

Dear Parents/Guardians:

We are writing to you to see if we may have your permission to involve you child in a research study that we are conducting. We are investigating different ways of translating provincial achievement tests from English to French. We are particularly interested in how English speaking students interpret achievement test items when they are written in English and how French Immersion and Francophone students interpret test items when they are translated from English into French. To determine this, we are interviewing a number of students and asking them to tell us how well they understand the items in Mathematics and Social Studies.

Your child may be randomly selected from a group of students for whom permission is granted for the interview. Each student who is selected will be asked to complete a short test containing either mathematics items or social studies items in English or in French if they are in a French Immersion program or if they are a Francophone. As students to be interviewed complete each item, they will be asked how they interpreted the question and whether they found any parts of the question confusing or helpful in answering the item. Each student will be tested individually. The total time required will be approximately 50 minutes (about one class period). If at any time during the interview your child decides he/she does not want to participate, he/she can immediately withdraw without prejudice.

All students will be interviewed at their school. The interviews will be audio-taped to ensure our results are accurate. The audio-tapes will be securely locked at the University of Alberta. Only we will have access to them. All responses provided by individual students will be completely confidential. The results will not be used for the students' school grades. Pseudonyms will be used in all reports; no student will be identified by name in any report.

The study has been approved by __________, principal of ___________, and by the Faculty of Education Research Ethics Board at the University of Alberta. A summary of the results will be available to school staff, students, and parents who are interested.

Thank you for considering our request. Please complete the form below and return it to your child's teacher. If you would like more information, please contact Jie Lin at 492-5427 or Dr. Todd Rogers at 492-3763 at the University of Alberta.

Sincerely,

W. Todd Rogers, Ph. D.	Jie Lin
Professor and Director	Doctoral student

Permission Form

May 10, 2005

________________________ (child's name) has my (our) permission to participate in the study of how students interpret Provincial achievement test items. I understand that the session will be audiotaped and that my son/daughter will not be identified in this study. I also realize that if my son/daughter decides not to participate at any time, he/she can withdraw from the interview without prejudice.

________________________ (child's name) does NOT have my (our) permission to participate in the study on how students' interpret math/social studies test items.

____

(Signature of parent(s)/guardian)

Please return this form, whether permission is granted or not, at your earliest convenience. Thank-you for responding.

Centre for Research in Applied Measurement and Evaluation
Faculty of Education
6-110 Education North · University of Alberta · Edmonton · Canada · T6G 2G5
Telephone: (780) 492-3762 · Fax: (780) 492-0001
www.ualberta.ca

Appendix J

Verbal Report Instructions for Social Studies

Hello ____________ (the student's name). My name is _____________.

How are you today?

Thank you for participating in this study. I am interested in how you understand the social studies questions that appear on a test. To find out about this, I am going to ask you to THINK ALOUD as you work through each question. By think aloud I mean that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you select an answer. It is important that you do not plan out or try to explain to me what you are thinking. Just act as if you are alone in the room speaking to yourself. Please try hard to talk about what you are thinking. If I notice that you have stopped talking, I will remind you to keep talking. Do you understand what I want you to do?

Note. Please don't talk to the student during THINK-ALOUD unless he/she stops talking for 5 seconds, remind him/her to "keep talking".

Please take you time, and answer the questions as best as you can. Afterwards, I am going to ask you a few questions about your understanding of the item. Before we start, I would like to remind you that I am not going to tell your teacher or your principal how you answered the questions. This study will not affect your mark in social studies. Do you have any questions before we begin?

We'll start with this practice question…

Note. Feel free to disrupt the subject in this warm-up exercise in order for him/her to get the idea of THINK ALOUD and how to answer the following questions.

Before each question

Please tell me what you are thinking as you answer this question. Please remember to say everything that is going through your mind.

After each question

Now there are a few questions that I would like to ask you about the item you just looked at:

1. Are there any words that you don't know in this question, including stimulus material (text in box, if applicable) and the question itself?

2. In your own words, could you tell me what you believe the question is asking? Imagine you are explaining it to a classmate.
    i. Possible follow-up probes: Why do you believe this?
3. Did you find any parts of the question confusing? If so,
    i. What parts did you find confusing?
    ii. Why are they confusing?

4. Did any find any parts of the question helpful in solving the problem? If so,
    i. What parts did you find helpful?
    ii. How did they help you solve the problem?

Possible probes when stimulus materials are involved and the information has NOT been provided previously:

i. Where did you find the information you need for this question?
ii. Describe in your own words what that part of the text means?

Further instruction for French-immersion students:

After a student has finished answering the above questions, you will show him/her the same item in English.

Now I would like you to read the English version of this item, and tell me if it means the same as the item in French. Feel free to mark things on the paper if you need to. Let's start with the stimulus material (if applicable).

Note. Ask Questions 5 and 6 separately for the stimulus material, the stem, and the options.

5. Do the two versions of the item mean exactly the same thing to you? On a scale--different, similar, identical--how would rate their comparability in meaning? How do the two versions differ in meaning? *(if "different" or "similar" was chosen)*

If they responded "different" or "similar", they were then asked:

6. Do you find any differences in wording? If so, where are they and how do the words differ?

At the end of the interview:

________________ (Student's name), thank you for your time and participation in this study. You have been really helpful to me. Thank you.

Appendix K

Observation Sheet

Date: ____________________ Interviewer name: ________________

School: ____________________ Language: ______________________

Student information:

	Name	Gender	Subject	Non-verbal behaviour*
1				
2				
3				
4				
5				
6				

* Please record any non-verbal behaviours of the student that may be relevant to the purpose of the study.

Appendix L

Confidentiality Agreement

This form was used for individuals hired to assist with interviews, and where necessary, to translate and transcribe the verbal report data conducted at the schools.

Project Title: Equivalence of Achievement Tests in English and French Developed Using the Simultaneous Test Development Approach

I, ______________________________, a bilingual interviewer have been hired to assist Jie Lin with her data collection procedures. I understand that I will be required to help conduct interviews, and translate and transcribe verbal reports.

I agree to:

1. keep all the research information shared with me confidential by not discussing or sharing the research information in any form or format (e.g., computer disks, tapes, transcripts, test items) with anyone other than Jie Lin or Dr. Todd Rogers.
2. keep all research information in any form or format (e.g., disks, tapes, transcripts) secure while it is in my possession.
3. return all research information in any form or format (e.g., disks, tapes, transcripts) to Jie Lin when I have completed the research tasks.
4. after consulting with Jie Lin erase or destroy all research information in any form or format regarding this research project that is not returnable to Jie Lin (e.g. information stored on a computer hard drive).

____________________	____________________	____________________
(print name)	(signature)	(date)
Jie Lin	____________________	____________________
(print name)	(signature)	(date)

This study has been reviewed and approved by the Faculties of Education and Extension Research Ethics Board at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Chair of the EE REB at (780) 492-3751.

Appendix M
Coding Scheme for Mathematics

Unknown words?	**Understanding?**	**Confusing?**	**Helpful?**	**Comparison**
No	Full	No	No	• S: Same / Identical • SI: similar • D: different • KA: Keep answer after reading English • CA: Change answer to … after reading English
Yes • List unknown words • how the students think they mean in brackets	Partial Explain why	Yes • Too much info • List the confusing part (specific to questions): • Note how the students think they mean in brackets	• AI: All the given info • OP: Options • List the helpful part (specific to questions)	• EC: English clearer / easier / flows better • FC: French clearer / easier / flows better • FE: More familiar with English terminology • FF: More familiar with French terminology • EN: English as a native language * More than one of the above may apply, use slash if so.
	No Explain why			If mentioned, list the specific words that are not equivalent in the two versions.

Appendix N
Coding scheme for Social Studies

Unknown words?	**Understanding?**	**Confusing?**	**Helpful?**	**Comparison**
No.	T (text): Full Q (question): Full	No	No	• S: Same / Identical • SI: similar • D: Different • KA: Keep answer after reading English • CA: Change answer to … after reading English
Yes. • List unknown words • Write how the students think they mean in brackets	T: Partial Explain why Q: Partial Explain why	Yes • List the confusing parts (specific to questions) • Note how the students think they mean in brackets	• OP: Options (list them) • List the helpful part or key words (specific to questions)	• EC: English clearer / easier / flows better • FC: French clearer / easier / flows better • FE: More familiar with English terminology • FF: More familiar with French terminology • EN: English as a native language * More than one of the above may apply, use slash if so.
	T: No/Little Explain why Q: No Explain why			If mentioned, list the specific words that are not equivalent in the two versions.

Appendix O

Coding Sheet for English Transcripts

Subject ________ Item No. ________

Name	Unknown words?	Understanding?	Confusing?	Helpful?	Remarks

Appendix P

Coding Sheet for French Transcripts

Subject ________ Item No. ________

Name	Unknown words?	Understanding?	Confusing?	Helpful?	Comparison	Remarks